University of Alberta

COGNITIVE-PSYCHOMETRIC MODELING OF THE MELAB READING ITEMS

by

Lingyun Gao ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Measurement, Evaluation, and Cognition

Department of Educational Psychology

Edmonton, Alberta
Fall 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-32958-0
Our file Notre référence
ISBN: 978-0-494-32958-0

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

A call has been issued within the measurement community to integrate cognitive psychology with assessment to inform test design and validation and to provide detailed diagnostic feedback (National Research Council, 2001). One approach to achieving this integration is to model item statistics, in particular item difficulty, in terms of the cognitive processes underlying item solving (Huff, 2003). To date, only a few models have been developed linking item statistics to the cognitive structure of test items or tasks, and these models are limited by the concepts and methods used. The purpose of this study was to use a cognitive-psychometric approach to model the reading items included in the MELAB, a large-scale high-stakes assessment of English as a second language.

The model was developed and tested through four stages. First, based on the review of the processes associated with reading and reading test taking by ESL learners, an initial cognitive model was hypothesized to underlie the MELAB reading test item performance. Then, this model was validated using a three-stage procedure: analyzing cognitive demands of the test items by raters, collecting students' verbal reports of the processes they used to arrive at the correct responses, and examining the relationship between the proposed cognitive processes and item difficulty estimates using the tree-based regression.

A review of the literature suggested that seven reading processes, four test-management processes, and three testwise processes should be included in the initial cognitive model underlying the ESL reading test item performance. Triangulation of the

three sources of evidence collected in the next three stages supported the conclusion that the seven reading processes and three test-management processes underlie successful performance on the MELAB reading test items.

This study demonstrated the value of using multiple sources of evidence to evaluate the performance of a cognitive model, and successfully demonstrated the union of cognitive psychology and assessment in the field of second/foreign language testing. While substantial evidence has been found in support of the cognitive model containing 10 processing components, the model warrants further research so that large-scale language testing programs can yield more meaningful results regarding examinees' reading abilities.

Acknowledgements

My first and deep appreciation goes to Dr. W. Todd Rogers, my academic and thesis supervisor, whose wisdom, insight, expertise, quick response, and unwavering support shaped my work and made my doctoral studies a quite rewarding experience. He had high standards and would not let go of my thesis until it was perfect. He had high expectations and was constantly pushing me to pursue higher professional goals. Overwhelmed as he was, he always made himself available for consultations even on weekends and holidays during the critical stages of my dissertation writing. He spent too many hours to count supervising this study from conceiving, executing, revising, to editing the numerous drafts of this thesis. His high intellectual caliber, enthusiasm, dedication, exemplary scholarship, and his compassion, unselfishness, and integrity will guide me through my professional career. I am truly blessed to have him as my mentor and my deep appreciation toward him is far beyond verbal expression!

I am deeply indebted to Dr. Linda M. Phillips, a distinguished researcher in the field of language and literacy, for her expertise and unconditional support and care given to me through the writing of this dissertation. She introduced me to the rich literature on reading and provided insightful feedback on my work along the way. Our meetings were usually short but extremely productive. She made me feel really confident about my initial cognitive model. Her extensive knowledge of reading, outstanding research she has conducted in this area, and exemplary scholarship have impacted my research in many positive ways and will make a unique contribution to my professional career.

I would like to express my sincere thanks to Dr. Jacqueline Leighton, who introduced me to the literature on cognitive psychology and inspired my dissertation research. She has been one of the greatest professors I have ever had. I was so impressed with her deep knowledge, clear organization, and thought-challenging questions when I took the classes Univariate Statistics, Educational Measurement, and Cognition with her. She guided me to think deep and hard about the conceptual things behind statistics and to think more about the global picture that I wanted to present in my study. This research would have been impossible without her prompt and insightful feedback and continued support.

I would like to express my heartfelt thanks to my external examiner, Dr. Lyle F. Bachman, for his thoughtfulness and the insightful comments that he made on my work. The expert contributions that this distinguished theorist and master practitioner has made to the language testing literature have had an enormous impact on my research.

I also owe thanks to Dr. Judy Cameron, who introduced me to learning and cognition. Her Cognition class was fascinating: She explained complex concepts in easy language and I never felt bored in her class! She was always there to support and encourage me during the last four years. She showed great interest in my work. She read my dissertation and reassured me of its good quality. I was so grateful that she was there to advise me on the ESL/EFL learning and testing aspects of my dissertation.

My sincere thanks also go to Dr. Connie Varnhagen, who has been so supportive of my work in various ways. Connie agreed to serve on my committee without any hesitation. I am grateful for her great interest in and very helpful comments on my work.

My heartfelt appreciation is extended to Dr. Mark Gierl, for his continued support through my doctoral program. His expertise in psychometrics, especially in item response theory and test equating, impacted my research in many positive ways and contributed substantially to my professional development. I owe thanks to Dr. Leslie Hayduk for explaining complex concepts of structural equation modeling in a non-daunting manner. I also own thanks to Dr. Julia Ellis, who introduced me to qualitative research methodology and made me feel so confident of myself! I would like to thank Dr. Carolyn Yewchuk, Dr. Derek Truscott, and Ms. Betty jo Werthmann for your help with my ethics application. I would like to thank the staff of our department and Josie Nebo at the Canadian Centre for Research on Literacy for your always providing me with full assistance whenever I needed it.

I would like to thank the English Language Institute (ELI) at the University of Michigan for sponsoring my dissertation research through the Spann Fellowship and providing me with the MELAB test materials and data for use in this study. I own thanks to the ELI staff for their interest in my research and support in whichever way they could. This study would never have been accomplished without their support.

My thanks also go to the raters and the students from the University of Alberta, who, despite their hectic schedules, tried to help me with the ratings and verbal reports for my dissertation data. I am particularly indebted to Jie Lin, Adele Tan, Changjiang Wang, and Yinggan Zheng, who spent countless hours assisting me with my data collection. Jie also provided insightful feedback on refining the rating instrument used in this study. I would like to thank all CRAMers for your comradeship and friendship. I enjoyed their brief but warm greetings and our occasional gatherings, which often lifted my day.

I would like to thank my parents, whose thoughts and unselfish love have been accompanying me whenever and wherever. They were always proud of me and never doubted I could make it. Their belief in me helped me persevere. They were always there when I needed them. They constantly encouraged me, tolerated my swaying mood, and hided their torture from disease and tiredness when I was in labor with my dissertation. I hope they can share the excitement of my success and forgive my being far away from them all these years.

My heartfelt thanks go to my beloved husband, Changjiang, for his continued support for me during this physically and emotionally trying period. He always believed in my capability and helped me in whatever way he could. He was always caring, loving, and forgiving, and always reminded me that there were other things in life that we need to pursue together. I would not be able to stand where I am today without him.

Lastly, I am very grateful to Albert, my little sweetie. His innocent love, smiling face, and energetic movements brought boundless happiness to my life. I hope he could forgive me for not watching and recording all his growing details during my trying period.

Table of Contents

CHAPTER 1: INTRODUCTION 1
The Michigan English Language Assessment Battery (MELAB) 5
Purpose and Research Questions 7
Significance of the Study 8
Definition of Terms 10
    ESL vs. EFL 10
    Skill, Strategy, Process, and Ability 10
    Construct 13
    Automatic vs. Controlled Processes 13
    Academic Reading 14
    Item Modeling 14
Organization of the Dissertation 14

CHAPTER 2: LITERATURE REVIEW 16
L2 Reading and Test-Taking Processes 16
    Information-Processing Perspectives on Reading 17
    Uniqueness of the L2 Reading 19
    L2 Reading Processes 21
        Knowledge of phonology, orthography, and vocabulary 21
        Knowledge of syntax 23
        Knowledge of textual features 23
        Background knowledge and subject matter/ topic knowledge 25
        Cognitive skills 26
        Cognitive and metacognitive strategies 27
        Purpose and context 28
    L2 Reading Test Taking Processes 28
The Assessment of L2 Reading 29
    A Theoretical Framework of Communicative Language Competence 30
    Assessing L2 Academic Reading Ability 34
        Constructs of the L2 academic reading ability 34
        The MELAB reading 35
    Variables that Affect the Difficulty of L2 Reading Test Items 37
Research into the Factors Affecting L2 Reading Test Item Difficulty 39
    Methods and Issues Concerning This Research 40
    Studies of Surface Task Characteristics and Item Performance 41
    Studies of Cognitive Demands of Test Items and Item Performance 44
    Processes in Task Performance Inferred from Verbal Reports 48
    Item Modeling with New Concepts and Methods 58
The Tree-based Regression (TBR) 60

- *A Description of the TBR* 60
- *Applications of the TBR to Item Modeling* 66

Literature Summary 70

CHAPTER 3: AN INITIAL COGNITIVE PROCESSING MODEL **74**

Cognitive Processes and Cognitive Variables 74
- *Reading Processes Assessed by the MELAB Reading* 74
  - *Word recognition* 74
  - *Vocabulary knowledge* 76
  - *Syntactic knowledge* 77
  - *Knowledge of discourse structure* 78
  - *Synthesis* 79
  - *Drawing inferences* 80
  - *Pragmatic knowledge* 81
- *Test Management Processes* 82
  - *Locating specific information requested by item* 82
  - *Matching information in question to information in text* 84
  - *Evaluating alternative choices* 85
  - *Using topical knowledge* 86
- *Testwise Processes* 87

A Summary of the Initial Model 88

CHAPTER 4: ITEM CODING BY RATERS **95**

Method 96
- *Raters* 96
- *Materials* 97
  - *MELAB test forms* 97
  - *Rating instrument* 97
- *Data Collection* 98
  - *Pilot study* 98
  - *Main study* 99
- *Data Analysis Procedures* 100
  - *Checking rating reliability* 100
  - *Reaching consensus ratings* 101

Results 101
- *Rating Reliability* 101
- *Rating Results* 103
- *Implications for the Model Revision* 108

CHAPTER 5: VERBAL REPORTS OF STUDENTS **111**

Method 111

- *Participants* 111
- *Materials* 112
  - *Background questionnaire* 112
  - *Verbal report practice tasks* 113
  - *MELAB test forms* 113
- *Data Collection* 113
  - *Pilot study* 113
  - *Main study* 115
- *Verbal Protocol Procedures* 117
  - *Coding the verbal report data* 117
  - *Coding reliability* 118
  - *Supplementary data analysis* 119
  - *Comparing verbal reports to ratings* 119

Results 120
- *Biographic Description* 120
- *Topic Familiarity* 123
- *Validating the Components of the Initial Model with the Verbal Reports* 127
- *Comparing the Item Feature Ratings to the Verbal Reports* 133
- *Implications for the Model Revision* 138

CHAPTER 6: ITEM DIFFICULTY MODELING USING TBR **142**

Method 142
- *Data* 142
- *Data Analysis Procedure* 143
  - *Data scoring and analysis of psychometric characteristics* 143
  - *Estimation of item parameters* 143
  - *TBR analyses* 144

Results 147
- *Psychometric Characteristics* 147
- *TBR Results* 151
  - *Test of the 10-component cognitive model (Form E)* 152
  - *Test of the 11-component cognitive model (Form E)* 157
  - *A comparison of the 10- and 11-component cognitive model (Form E)* 162
  - *Test of the 10-component cognitive model (Form F)* 163
  - *Test of the 11-component cognitive model (Form F)* 169
  - *A comparison of the 10- and 11-component cognitive model (Form F)* 174
  - *A comparison of the 10- and 11-component model across the test forms* 176

CHAPTER 7: DISCUSSION AND CONCLUSION **183**

Summary of Methods and Discussion of the Findings 183
- *Development of the Initial Cognitive Model* 183

Cognitive Analysis of the Items by Raters.. 185
Student Protocols and Analysis 187
Item Difficulty Modelling using Tree-Based Regression 189
Limitations.. 192
Conclusion.. 193
Implications for Educational Practice.. 194
Directions for Future Research 196

References 200

Appendix A: Recruiting Letter and Letter of Information for Raters 220
Appendix B: Rating Instrument 222
Appendix C: Consent Form for the Raters 227
Appendix D: Procedures for Rater Training and MELAB Reading Item Coding ... 228
Appendix E: Summary of Item Ratings by the Three Raters 230
Appendix F: Recruiting Letter and Letter of Information for Verbal Participants ... 234
Appendix G: Procedures and Instructions for Verbal Report Sessions 236
Appendix H: Practice Tasks for the Verbal Reporting Sessions 238
Appendix I: Consent Form for the Verbal Repot Participants 239
Appendix J: Confidentiality Agreement 240
Appendix K: The Initial Tree for the 11-component cognitive model (Form E)......241
Appendix L: The Initial Tree for the 10-Component Cognitive Model (Form F).....242
Appendix M: The Initial Tree for the 11-Component Cognitive Model (Form F)....243
Appendix N: Correlation Matrix (Form E).. 244
Appendix O: Correlation Matrix (Form F)..245

List of Tables

Table 1: Cognitive Processes Used by L2 Readers When Answering the Multiple-Choice Reading Test Items 90

Table 2: Variance Components and Generalizability Coefficient from the G-Study and D-Study 102

Table 3: Final Set of Consensus Ratings 104

Table 4: Summary of Consensus Item Ratings for Form E and Form F 105

Table 5: Characteristics of the Verbal Report Participants 121

Table 6: Descriptive Statistics for the Verbal Report Participants.................. 123

Table 7: Distributions of MELAB Topic Familiarity by Test Form 124

Table 8: Frequencies of Cognitive Processes Used for the Correct Responses (Form E) 128

Table 9: Frequencies of Cognitive Processes Used for the Correct Responses (Form F) 129

Table 10: Comparisons between Verbal Data and Rating Data (Form E) 134

Table 11: Comparisons between Verbal Data and Rating Data (Form F) 135

Table 12: Frequencies of Occurrence within Each Levels of Consistency (Form E)... 136

Table 13: Frequencies of Occurrence within Each Levels of Consistency (Form F) 137

Table 14: A Comparison of the Verbal and Rater Results Regarding the Model

Components 140

Table 15: Descriptive Statistics and Reliability for the Two Forms 148

Table 16: Item Difficulty Parameter Estimates for All Items 151

Table 17: The Final Trees for the Two Cognitive Models and Two Forms 177

List of Figures

Figure 1. Components of communicative language ability 31

Figure 2. Components of language competence 32

Figure 3. Illustration of the TBR splitting rules 63

Figure 4. Illustration of the Tree-Based Regression Analysis 64

Figure 5. Model Development and Testing Procedures 72

Figure 6. Frequency of the cognitive processes used by the Forms E and F students 131

Figure 7. Histogram of total scores on Form E 149

Figure 8. Histogram of total scores on Form F 149

Figure 9. Test characteristic curves for Form E and Form F 150

Figure 10. Tree diagram for the cognitive model with 10 components (Form E) ... 153

Figure 11. Tree diagram for the cognitive model with 11 components (Form E) ... 159

Figure 12. Tree diagram for the cognitive model with 10 components (Form F) ... 165

Figure 13. Tree diagram for the cognitive model with 11 components (Form F) 170

CHAPTER 1: INTRODUCTION

Large-scale assessments are widely used for a variety of purposes such as measuring student achievement and matching students to appropriate instructional programs. The ultimate goal of testing is to enhance learning (National Research Council [NRC], 2001). In most cases, assessment results provide general information about where a student stands relative to others (e.g., that the student scored at the 90th percentile) or how the student has performed on the test (e.g., that the student correctly answered 50 out of 100 test items). A serious limitation with current large-scale assessments, however, is that they are incapable of providing more detailed information that can be incorporated into the learning system, such as where the students' strengths and weaknesses are, why they perform poorly, and how to improve their performance (Alderson, 2005a; Embretson & Gorin, 2001; NRC, 2001). For example, for a student who performed poorly on a reading test, a numeric score of 50 out of a total of 100 points does not tell whether this student lacks word knowledge to understand the text or lacks strategic competence to answer the questions. It is frequently found that students who achieve the same score have different levels of understanding and that students who correctly answer many questions lack the tested skills (e.g., Gao, 2002; Katz & Lautenschlager, 1995; 2001). As a result, tests provide very limited feedback to students, educators, administrators, and other stakeholders (Alderson, 2005b; Strong-Krause, 2001). In the case of large-scale tests used primarily for admission purposes, such as the *Test of English as a Foreign Language* (TOEFL), the validity of test scores as an admission tool is frequently questioned and the usefulness of the tests as a learning tool is compromised (Huff, 2003).

In the last two decades or so, a call has been issued for large-scale assessments to yield more meaningful information to benefit learning and instruction (Mislevy, 1996; Nichols, 1994). Developments from two fields, cognitive psychology and measurement, provide a scientific basis to accomplish this goal (NRC, 2001). Cognitive psychology is the study of how people learn, organize, store, and use knowledge (Sternberg & Ben-Zeev, 2001). Conceptually, cognitive research suggests not only what aspects of learning are important to measure, but also how test tasks can be designed to provide evidence for the desired inferences (Messick, 1995; NRC, 2001). Contemporary cognitive theories acknowledge the active role of learners in constructing their understanding and emphasize the processes in which the learners represent, organize, and process knowledge (Sternberg & Ben-Zeev, 2001). Accordingly, the aspects of learning that a test should measure have included cognitive processing skills or strategies, such as reasoning (e.g., deduction, inference), problem-solving (e.g., using prior knowledge and metacognition), and decision-making (e.g., confirmation and disconfirmation of answer choices). Methodologically, the methods that cognitive psychologists use to understand human thought have been applied in testing research to investigate examinees' cognitive processes. Some of the most useful methods include the study of reaction times, protocol analysis (i.e., the analysis of people's subjective reports as they solve problems), and computer simulations. The cognitive perspective of testing helps diagnose an examinee's strengths and weaknesses in thinking and problem solving, and thus provides richer feedback to score users than a single numeric score or percentile rank (NRC, 2001).

Measurement models are statistical models used to draw inferences about proficiencies based on the information obtained from test tasks (NRC, 2001). Over the last several decades, item response theory (IRT) models have had a major impact on testing. Assuming that performance of an examinee on a test item can be explained by a set of latent traits or abilities, IRT models specify the relationship between item performance and the abilities underlying item performance (Hambleton & Swaminathan, 1985; Lord & Novick, 1968). Another assumption common to the IRT models is that they are conditional upon abilities, whereby examinees' responses to different items are statistically unrelated (i.e., assumption of local independence). The most commonly used IRT models are the one-, two-, and three-parameter logistic models (van der Linden & Hambleton, 1997).

The IRT models have advantages over the classical test theory (CTT) model. The most desirable feature of the IRT models is the invariance of item and ability parameters. That is, item indices do not depend on the ability distribution of the examinees being measured and the ability of the examinees does not depend on the set of items administered (Hambleton, Swaminathan, & Rogers, 1991). Despite the advantages of the IRT, without the linkage to contemporary theories of cognition, the IRT models cannot yield rich inferences about examinees' knowledge and skills (NRC, 2001). As Embretson (1993) argued, "they (traditional IRT models) have little connection with the concerns of cognitive theory about processes, strategies, and knowledge structures that underlie item solving" (p. 125) and lack the capacity to interpret more complex forms of evidence derived from student performance.

To overcome this limitation, several cognitive IRT models have been proposed that capture examinees' cognitive processes based on patterns of responses to test items, such as the logistic latent trait model (LLTM) (Fischer, 1973) and the multiple latent trait model (MLTM) (Whitely, 1980). More recently, in response to the call for an integration of cognitive psychology and measurement (Embretson & Gorin, 2001; Leighton, 2004; NRC, 2001), a variety of measurement models have emerged that incorporate cognitive elements, such as the tree-based regression model (Sheehan, 1997), rule-space model (Tatsuoka, 1995), attribute hierarchical model (Leighton, Gierl, & Hunka, 2004), and Bayes inference networks (Mislevy, Almond, Yan, & Steinberg, 1999). While these models may differ in the way in which cognitive information is used, they have two common elements. First, it is generally assumed that the ability to answer particular sorts of test questions involves a variety of cognitive components, and, because of this, these models have mainly been applied to mathematics (e.g., Ewing & Huff, 2004) and reading (e.g., Van Essen, 2001), where multiple cognitive components are assumed to be existing and related to the ability to answer a test question (Dibello, Stout, & Roussos, 1995; Gorin, 2002). Second, the statistical strategy for incorporating cognitive elements in these models is to add new parameters that describe the cognitive structures required by item solution (NRC, 2001).

The emergence of these complex measurement models makes it possible to incorporate what cognitive psychologists consider important to pursue: to provide substantially rich information on the cognitive demands of test tasks and to reveal meaningful cognitive processes of students (NRC, 2001). Nevertheless, due to technical complexity and lack of cooperation among cognitive psychologists, measurement specialists, educators, and domain experts, many of these models have not been widely applied to testing practice (NRC, 2001). This is true in the field of second/foreign language testing. Over the last decade, Educational Testing Service (ETS) has taken the lead to apply new measurement models to the TOEFL testing program (Buck, Tatsuoka, & Kostin, 1997; Huff, 2003) and launched a new generation of the TOEFL in 2005, incorporating principles of cognitive psychology (Mislevy, Steinberg, & Almond, 2002). However, the cognitive measurement models have been applied to few other language testing programs. Much work is required to link measurement models to critical features of cognitive models specific to a substantive testing context and to observations that reveal meaningful cognitive processes in a particular domain. This process requires not only statistical modeling but also substantive theories of the target domain (NRC, 2001).

The Michigan English Language Assessment Battery (MELAB)

Developed by the English Language Institute of the University of Michigan (ELI-UM), the MELAB is administered regularly at the 85 test centers in the United States and Canada following uniform procedures. According to the MELAB technical manual (ELI, 2003), the MELAB is designed to assess advanced-level English language competence of adult non-native speakers of English, who will need to use English for academic study in a university setting. The theoretical framework for developing the MELAB is closely related to the model of the communicative language ability construct proposed by Bachman (1990) and later revised by Bachman and Palmer (1996). According to this model, communicative language ability consists of " both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use" (Bachman, 1990, p. 84).

The MELAB is used primarily for the purpose of higher education admission. The assessment results are widely accepted as evidence of English competence, as an alternative to the TOEFL, by educational institutions in the United States, Canada, United Kingdom, and other countries where English is the language of instruction. In addition, many scholarship programs, government agencies, and licensing/certification agencies use the MELAB results to make high stakes decisions about educational and employment opportunities (ELI, 2003).

The MELAB consists of three parts. Part 1 is a written composition. Part 2 is a listening test containing 50 multiple-choice items. Part 3 is a written test containing 100 multiple-choice items (30 grammar, 20 cloze, 30 vocabulary, and 20 reading items). The speaking test is optional and not included in every MELAB administration. Composition and speaking tests are scored by trained raters using rating scales. Answer sheets for Parts 2 and 3 are computer scanned and raw scores (i.e., the total number of test items answered correctly) are converted to scale scores. The MELAB reports a score for each part and the final score, which is the average of the scores on the three parts. Scores on the speaking test are reported separately. As the manual (ELI, 2003) explains, part scores are intended to report examinees' competence in separate language skill areas, while the final score is to report examinees' general competence in English.

Current MELAB score reporting provides some information on examinees' English proficiency and the use of rating scales for scoring compositions and speaking tests describes examinees' proficiency in writing and speaking to some extent. However, a numeric score for Parts 2 and 3 provides very limited information to examinees, admission officers, and other stakeholders regarding examinees' strengths and weaknesses in listening and, especially, in reading where a sub-score is lacking. Reading is a major part of language acquisition and language use activity in everyday life (Grabe & Stoller, 2002). In the context of using English as a second or foreign language for academic purposes, reading tends to be the single most important language use activity and language skill that non-native English speakers need for academic activities (Carr, 2003; Cheng, 2003). Thus, understanding the nature of reading in a second or foreign language and how to assess it on large-scale high-stakes tests have become primary concerns for language researchers and testers (Alderson, 2000; 2005a; 2005b; Bernhardt, 2003; Cohen & Upton, 2006).

Purpose and Research Questions

The purpose of this study was to model the cognitive processes underlying the MELAB reading test item performance using a cognitive-psychometric approach. The specific research questions addressed were:

1. What components should be included in the initial cognitive processing model for the MELAB reading test items?
2. What cognitive processes are required to correctly answer the MELAB reading test items?
3. What cognitive processes are actually used by examinees when they correctly answer the MELAB reading test items? How are they related to the findings in response to question 2?
4. To what extent do the cognitive processes identified in response to questions 2 and 3 explain the item difficulty parameter estimates?

To develop this model, theoretical information of the reading processes in a second or foreign language was considered and operational definitions of the cognitive processes were developed from the reading items included in the MELAB. Next, the hypothesized model was validated using a three-stage procedure. Specifically, the first stage of the model validation involved identifying the cognition demands of the MELAB reading items by judges. In the second stage of the model validation, the cognitive processes that examinees actually used when they correctly answered the MELAB reading items were investigated. In the last stage of the model validation, the proposed cognitive processes were validated through empirical studying of the objective performance on the MELAB reading items, using an advanced measurement model called the tree-based regression.

Significance of the Study

The results of this study will contribute to the theory, methodology, and practice of testing ESL/EFL reading for academic purposes, and to the field of second or foreign language testing in general. Theoretically, the review and analysis of the cognitive processes involved in correctly answering the MELAB reading items will increase our understanding of the nature of ESL/EFL reading and the constructs of ESL/EFL reading ability. Moreover, modeling the cognitive processes underlying the MELAB reading item performance using a novel measurement model, the tree-based regression, will contribute to our understanding of the relationship between item features and item difficulty.

Methodologically, the resulting model will link the cognitive theories in the domain of ESL/EFL reading to a measurement model and be supported by theory and empirical data. The embracement of a theoretical review of the ESL/EFL reading processes and substantive analysis of the reading test items, which is lacking in current research on the MELAB, will make possible theory-based test development and score interpretations (Messick, 1995). Moreover, embarking on qualitative analysis of the cognitive processes required or used to solve the items and quantitative analysis of objective performance will demonstrate a unified procedure to validate a theoretical model (Bachman, Kunnan, Vanniarajan, & Lynch, 1988; VanderVeen, 2004). Finally, the use of verbal protocol analysis to investigate examinees' cognitive processes and tree-based regression to model the cognitive processes underlying item performance will contribute to the progress of new methods for language testing research.

Practically, the resulting model of cognitive processes underlying performance on the MELAB reading items will have implications for the construct validity of the

MELAB reading as a measure of ESL/EFL reading proficiency required for college-level academic study (Embretson, 1998; Gorin, 2002; Huff, 2003). In addition, the results of this study may guide test developers to design cognitively-based reading items (Enright, Morley, & Sheehan, 2002). As Gitomer and Rock (1993) suggest, "improved test design consists of building items that are constructed on the basis of an underlying theory of problem-solving performance" (p. 265). Most importantly, the results of this study can be used to develop descriptive score reports and lay a foundation for the MELAB as a diagnostic measure. In this manner, large-scale language testing programs will be able to provide more meaningful feedback to score users about examinees' strengths and weaknesses in particular skills, suggest areas for improvement, and target instruction to individual needs (DiBello & Crone, 2001; Huff, 2003; Sheehan, 1997; Wainer, Sheehan, & Wang, 2000).

Definition of Terms

ESL vs. EFL: ESL is used where the context facilitates English learning or acquisition (e.g., a student whose first language is Mandarin learning English in Canada). In contrast, EFL is used where the context does not facilitate English learning or acquisition (e.g., a student whose first language is Mandarin learning English in China). In this study, both ESL and EFL are termed as L2 unless otherwise mentioned.

Skill, Strategy, Process, and Ability: These terms are not clearly distinguished in the literature and are often used interchangeably. To understand the term *cognitive processes* used in this study, distinctions between these terms are clarified as follows.

Skills are automatic information-processing techniques that a reader uses unconsciously while interacting with written texts or taking reading tests (Paris, Wasik, & Turner, 1991). According to Urquhart and Weir (1998), skills have three features. First, skills are cognitive and part of the generalized process of reading and reading test-taking. Second, skills are deployed automatically and unconsciously. Third, skills are text or task-oriented. That is, skills focus on characteristics inherent in the text or test task itself without taking into account the readers (e.g., how readers might process the text or how familiar they might be with the text topic).

Strategies are conscious efforts that a reader deliberately makes to construct meaning, solve problems in understanding, and answer questions on reading tests (Alderson, 2000). According to Urquhart and Weir (1998), strategies are essentially problem solving, which involves selection and efficiency, and have several features. First, like skills, strategies are cognitive and part of the generalized process of reading or test-taking. However, strategies are purposeful and goal-oriented. Second, strategies are adopted consciously, describable, and teachable. Third, strategies are reader-oriented. Cheng (2003) points out that there is no clear-cut distinction between skills and strategies. Skills can become strategies when they are used in a deliberate and purposeful manner and strategies can become skills when conscious actions become automatic through practice or training.

Process is used in relation to product, the result of that process. The literature on L2 reading assessment generally categorizes the processes used on reading tests into reading processes and test-taking processes. For example, Farr, Pritchard, and Smitten (1990) define reading processes as those used by readers when they read the passages before turning to the test items, and test-taking processes as those used by readers once they start reading the test items. The processes of reading and reading test-taking are commonly viewed as cognitive and part of a general process of reasoning and problem solving (Alderson, 2000; Bernhardt, 2000; Cohen & Upton, 2006). In the present study, the term *cognitive process* is used to refer to the process used by L2 readers to arrive at answers to the multiple-choice reading items included in the MELAB. This process is viewed as cognitive in nature and involving multiple components related to the use of knowledge, skills, and problem-solving strategies in the context of taking academic reading tests (Cohen & Upton, 2006; Urquart & Weir, 1998). The term *cognitive process* is used in a broad sense in this study, involving the reading processes assessed by the MELAB reading tests (e.g., identifying the structure of the text, identifying word meanings from context), test management processes (e.g., eliminating the options) and testwiseness (e.g., using educated guesses), all of which can be involved in answering the multiple-choice reading test items.

Ability refers to interpretations and inferences made based on observable performance and is observed through assessment tasks requiring those aspects of the process that testers consider important to measure (NRC, 2001). In the reading literature, ability is commonly viewed as involving distinct aspects, such as ability to make inferences and ability to recognize the main idea, though issues about divisibility of reading ability has not been completely settled (Carr, 2003). This study acknowledges the existence of definable aspects of reading ability underlying performance and operationalises them in light of Bachman's (1990) theoretical framework of communicative language ability. That is, ability in this study involves abilities to apply knowledge, skills, and strategies in a meaningful context.

Construct: Construct refers to what a test is intended to measure and represents examinees' latent traits inferred from observable performance (NRC, 2001). Constructs are commonly defined in terms of a theory of the ability to be tested. In this study, the constructs are the L2 reading abilities required for academic study at the college level, and are defined in light of Bachman's (1990) theory of communicative language ability.

Automatic vs. Controlled Processes: The process of reading and taking reading tests involves a great deal of automatic and conscious mental activities (Alderson, 2000). Automatic processes are beyond conscious control and are often referred to as "skill" (Williams & Moran, 1989). In this study, recognizing the meaning of basic vocabulary, using knowledge of syntax and text features to understand simple structures of sentences and texts are considered as automatic processes for the target examinees of the MELAB, that is, the advanced-level L2 learners having acquired basic English language knowledge and competence.

Controlled processes refer to those conscious processes during which examinees deliberately and purposefully exert an active control over their processes of reading and test-taking through the use of strategies (Williams & Moran, 1989). In the present study, controlled processes refers to the cognitive strategies that examinees consciously use to interact with the text and to answer the test items, such as synthesizing, drawing inferences, using prior knowledge, scanning through the print for the required information, and eliminating the impossible options to select the one that best fits. In addition, when the reading or the route to solve an item is unclear or ambiguous, readers may use the strategies such as guessing the meaning of an unknown word using context clues and determining how different parts of the text function in the discourse.

Academic Reading: refers to reading texts in a specific subject area for the purpose of learning in that area. In this study, it is conceptualized as a meaning-based activity that is closely related to tasks, texts, readers' linguistic knowledge, subject matter knowledge, strategic competence, reader purpose, and the context of that purpose (Douglas, 2000).

Item Modeling: refers to "the process of explaining the variance of an item's psychometric properties, such as difficulty or discrimination, with the features of the item, such as its content specifications, format, or the cognitive skill(s) and/or process(es) required to solve the item" (Huff, 2003, p. 19). In this study, it refers to mathematical modeling of an item's difficulty in terms of the cognitive processes used to correctly answer the item.

Organization of the Dissertation

This dissertation is organized into seven chapters. Chapter 1 introduces the context and rationale of the present study, highlights the purpose and research questions, and provides operational definitions of the terms used in this dissertation. Chapter 2 is a review of the literature, including (a) an overview of the theoretical positions related to the L2 reading processes, (b) a review of the theoretical framework and ability constructs pertaining to the L2 reading assessment, (c) a critical review of the empirical studies of the factors that affect item performance on the L2 reading tests, and (d) a discussion of the key measurement model used in this study, the tree-based regression. Chapter 3 outlines an initial model of cognitive processing following an analysis of the literature. Chapter 4 outlines the methods used and presents the results obtained for the first stage of the model validation – the identification of cognitive demands of the test items included in two forms of the MELAB reading test. Chapter 5 outlines the method used and presents the results for the second stage of the model validation – verbal protocols from the Mandarin-speaking students, who represent one of the largest language groups of the MELAB candidates. Chapter 6 presents the method and results for the third stage of the model validation – relating the proposed cognitive processes to the item difficulty estimates using the tree-based regression. Chapter 7 summarizes the methods and findings for each stage, discusses the limitations, and presents the conclusions of this study. The chapter concludes with a discussion of the implications for educational practices and directions for future research.

CHAPTER 2: LITERATURE REVIEW

This chapter presents the theoretical framework and empirical evidence relevant to modeling the MELAB reading items. The purpose of this literature review is threefold: (1) to provide theoretical support for the development of the cognitive processing model underlying the MELAB reading item performance; (2) to define a list of cognitive processes that would guide the analyses of the MELAB reading items and the verbal report data, and (3) to frame a new approach to modeling item performance. The review is organized into four main sections. The first section discusses the L2 reading and test-taking processes. The second section reviews the literature pertaining to the L2 reading assessment. The third section critically reviews empirical studies of the factors affecting the L2 reading test item performance, mainly multiple-choice test item difficulty, which have specific relevance to the current study. The last section reviews the psychometric literature related to the quantitative methodology used for this study, the tree-based regression.

L2 Reading and Test-Taking Processes

Understanding the processes of reading and test-taking used by examinees on tests of reading in a second or foreign language is critical to understanding and assessing L2 reading (Cohen & Upton, 2006). This section begins with a review of the theoretical positions on reading as a general information-processing process. Then, it discusses the uniqueness of the L2 reading. Finally, it reviews the literature on the L2 reading and test-taking processes.

Information-Processing Perspectives on Reading

Over the last couple of decades, the shift in psychology from a behavioral to a cognitive orientation has impacted enormously the understanding of reading. Bottom-up processing is an immediate left-to-right processing of the input data through a series of discrete stages (Ruddell, Ruddell, & Singer, 1994). Early theories viewed reading as bottom-up processing in which a reader passively and sequentially decoded meanings from letters, words, and sentences (e.g., Anderson, 1972; Bormuth, 1969; LaBerge & Samuels, 1974). The reading processes were considered to be completely under the control of the text and had little to do with the information possessed by a reader or the context of discourse (Perfetti, 1995).

Opposite to bottom-up processing, top-down processing refers to the way of information processing in which readers approach the text with their already-existing knowledge and then work down to the text (Hudson, 1998). The top-down view of reading emphasizes readers' contribution over the textual information. Two representative examples of top-down processing models are psycholinguistic models (e.g., Goodman, 1967, Smith, 1971) and schema-theoretic models (e.g., Carrell, 1983a; 1983b). Psycholinguistic models stress the interaction between language and thought, especially readers' inferential abilities, and describe reading as an active, purposeful, and selective process (Smith, 2004). According to psycholinguistic models, readers predict or guess the meaning based on "minimal textual information, and maximum use of existing, activated, knowledge" (Alderson, 2000, p. 17).

Schema-theoretic models describe the reading process through the activation of schemata (i.e., networks of information organized in memory) and stress the centrality of readers' language and content knowledge. During the process of reading, readers apply their schemata to the text, confirm and disconfirm, and map the incoming information from the printed text onto their previously formed knowledge structures to create meaning (Ganzter, 1996; Hudson, 1998). Schema theory is valued at attempting to explain the integration of the new information with the old, but it fails to explain how completely new information is processed (Alderson, 2000). Critics of schema theory point out that it lacks strong supporting evidence and is not scientifically testable. In addition, it does not lead to an explicit account of reading processes due to the vague definition of schema, oversimplification of the memory retrieval and storage processes, and elision of readers' intentionality (Phillips & Norris, 2002). Carver (1992a) argues that schema theory applies only when reading texts are relatively hard, such as the situation in which college-level students read academic texts.

More recent theories of reading stress the simultaneous interaction between bottom-up and top-down processing (e.g., Johnston, 1984; Rumelhart, 1977, 1980; Stanovich, 1980, 2000). According to the interactive theories, readers' multiple sources of knowledge (e.g., linguistic knowledge, world knowledge) interact continuously and simultaneously with text. Current reading theories acknowledge the interactive nature of processing, and emphasize the importance of purpose and context to fluent reading (e.g., Alderson, 2000; Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, & Schedl, 2000;

Hudson, 1998). As Butcher and Kintsch (2003) note, reading is the interaction among a variety of top-down and bottom-up processes, during which readers' knowledge, cognitive skills, strategy use, and purpose of reading are crucial and must be taken into account when modeling text processing.

Uniqueness of the L2 Reading

Research of L2 reading has long been dominated by the view that the L2 reading is fundamentally the same as the first language (L1) reading (Phakiti, 2003). Recently, however, there have been doubts and attacks on the overgeneralization that the L2 and L1 reading are the same (e.g., Bernhardt, 2003; Enright et al., 2000). These researchers argue that no existing theories, models, or views of L2 reading are capable of explaining L2 readers' reading process. They are concerned that the lack of thorough investigation of the L2 reading may influence the interpretation of the L2 reading constructs. To understand the processes of reading and test-taking on tests of reading in a second or foreign language, the uniqueness of the L2 reading is discussed briefly below.

First, L2 and L1 readers have different knowledge structures. On the one hand, L2 readers have limited knowledge of the second or foreign language, which may constrain their processing in L2. Unlike L1 readers, who have acquired basic language competence prior to reading instruction and are continuously exposed to written symbols in their culture, L2 readers read the second or foreign language before attaining adequate oral proficiency and have to deal with the materials not in their initial cultural environment (Enright et al., 2000). On the other hand, L2 readers have additional sources of knowledge related to their native language, culture, and rhetoric to overcome their linguistic limitations and facilitate meaning construction (Gantzer, 1996; Skehan, 1998). Research has shown that when L2 readers reach a threshold in reading proficiency, their general and domain-specific knowledge has a facilitating effect on text comprehension and recall (Clapham, 1996). For example, for the L2 readers who have acquired basic reading proficiency, the topics that are concrete, realistic, culturally more familiar, and closer to a reader's preexisting knowledge tend to be easier to understand than the topics that are abstract, arcane, and culturally or contextually less familiar.

Second, cognitive skills and strategies (e.g., deduction, inference, and monitoring) play an important part in L2 reading (Alderson, 2000; Koda, 2005; Skehan, 1998). Unlike reading in the native language, which begins at an early age, reading in a second or foreign language usually begins when L2 readers are teenagers or adults. Hence, L2 readers' cognitive resources (e.g., subject knowledge and problem-solving strategies) developed from reading in their native language enable them to fill in the comprehension gap when reading in L2 (Urquhart & Weir, 1998). Koda (2005) points out that with the development of cognitive abilities, older and more advanced students would be able to apply their L1 reading skills and strategies to L2 reading tasks and monitor their L2 reading process to judge whether a particular strategy is effective. Moreover, research studies repeatedly show that L2 readers depend heavily on their prior knowledge when they read in L2 (e.g., Gantzer, 1996; Gao, 2002; Johnston, 1984). In the context of this study, the target examinees of the MELAB are advanced-level adult non-native speakers of English who will need to use English for academic studies at the college level. They typically learned to read English after acquiring their L1 reading proficiency, have mastered basic English language knowledge, and are capable of using cognitive skills and cognitive and metacognitive strategies to cope with the L2 reading tasks.

L2 Reading Processes

The conceptualization of reading has been evolving over the years; so has the understanding of the L2 reading processes. Current views of L2 reading emphasize the important contribution of bottom-up processing to fluent reading and acknowledge readers' active role during reading (Alderson, 2000; Cohen, 2005; Urquhart & Weir, 1998). Consistently, current models of the L2 reading process have generally included language knowledge, general or domain-specific background knowledge, cognitive skills, and cognitive and metacognitive strategies. Language knowledge, often referred to as "bottom-up processing" or "lower-level text-based processing", consists of a number of relatively independent components, such as the knowledge of phonology, orthography, vocabulary, syntax, and text features. Background knowledge, often referred to as "top-down processing" or "higher-level knowledge-based processing", represents the knowledge that a reader brings to a text (Cohen & Upton, 2006). Major components in current models of L2 reading process are discussed in this section.

Knowledge of phonology, orthography, and vocabulary. Word recognition and vocabulary knowledge have been considered central to fluent reading in current models of the reading process of skilled adult L2 readers (e.g., Alderson, 2000; Carrell & Grabe, 2002; Koda, 2005; Hudson, 1996; Urquhart & Weir, 1998). Word recognition refers to the process of recognizing strings of letters that form words in print and being able to rapidly identify meanings from visual input (Rayner & Pollatsek, 1989). Two word recognition processes in reading have been distinguished: the phonological and the orthographic. The former requires "awareness of phoneme-grapheme correspondences and the word's phonological structure", while the latter requires memory for specific spelling patterns (Alderson, 2000, p. 344). In the recent reading literature, there is a growing consensus that all reading involves phonological and orthographic processing, which compose the process of identifying the word meanings (Alderson, 2000, Stanovich, 2000). Unlike skilled adult L1 readers who are generally assumed to have phonological access to the lexicon and are familiar with the script, L2 readers encounter words that they have not heard pronounced and scripts that they are not familiar with in many cases (Urquhart & Weir, 1998). Hence, L2 readers are expected to experience greater difficulty in processing letters in a word and identifying word meanings, which may affect their reading in L2 (Alderson, 2000).

In addition to word recognition, research studies of L2 reading have consistently shown that vocabulary knowledge is crucial to text comprehension and reading performance (e.g., Carrell & Grabe, 2002; Qian, 1999). Unlike skilled adult L1 readers for whom the words encountered are normally in their lexicon, L2 readers have to handle unfamiliar vocabulary (Urquhart & Weir, 1998). In the context of academic reading, where large amounts of academic texts need to be processed, efficient word processing is extremely important. Inefficient word recognition and insufficient vocabulary knowledge would likely result in inefficient academic reading (Hudson, 1996).

Knowledge of syntax. In addition to word processing, readers must process syntax in order to impose meaning on the recognized words (Urquhart & Weir, 1998). Syntax is the component of a grammar that determines the way in which words are combined to form phrases and sentences (Radford, 2004). In L2 reading, syntax knowledge is crucial for successful text processing, and syntactic processing has been included in many models of the L2 reading processes (e.g., Carrell & Grabe, 2002; Hudson, 1996; Koda, 2005).

Knowledge of textual features. Readers' knowledge of textual features, such as cohesion and text structure, has long been considered important in text processing (Alderson, 2000; Koda, 2005) and critical to successful L2 academic reading (Hudson, 1996). Cohesion refers to "the connections between sentences", which are furnished by pronouns that have antecedents in previous sentences, adverbial connections, known information, and knowledge shared by the reader (Kolln, 1999, p. 271). Cohesion occurs where understanding new information in the text depends on understanding the already available information (Hudson, 1996). Frequently used cohesive devices include reference, substitution, and ellipsis to replace previously occurring parts of the text, and conjunction to indicate "the pragmatic relationship between two text utterances or blocks" (Urquhart & Weir, 1998, p.74).

According to Thompson (2004), reference is the set of grammatical resources used to repeat something mentioned in the previous text (e.g., the pronoun "it") or signal something not yet mentioned in the text (e.g., the non-definite article "A" in the sentence "They came again into their bedroom. *A* large bed had been left in it"). Substitution refers to the use of a linguistic token to replace the repetitive wording (e.g., "Do you like it? I think *so*"). Ellipsis is the set of grammatical resources used to avoid the repetition of a previous clause (e.g., "How old is he? Two years old"). Conjunction refers to the combination of any two textual elements into a coherent unit signaled by conjunctions (e.g., however, by the way, thus).

Thompson (2004) further distinguishes cohesion and coherence, two terms often used interchangeably in describing textual features. According to Thompson, cohesion refers to the linguistic devices used to signal the coherence of the text. Textual features that serve a cohesive function can be identified. However, coherence is "in the mind of writer and reader", which "cannot be identified or quantified in the same way as cohesion" (p. 179). Coherence of a text depends on not only cohesive devices but also text structure and organization pattern, that is, how the sentences and paragraphs relate to each other and "how the relationship between ideas are signaled or not signaled" (Alderson, 2000, p. 67). Examples of text structures include cause/effect, general/specific, problem/solution, comparison/contrast, and the use of illustration, classification, and topic sentence. Research has shown that the internal logic of text structures (strong or weak), organized patterns (tight or loose), and location of information within text (earlier or later) affect understanding (e.g., Carrell, 1984, 1985; Roller, 1990; Hudson, 1996). Coherent texts contribute to understanding, while ambiguous references, indistinct relationships between elements in the text, and inclusion of irrelevant ideas or events hinder comprehension (Alderson, 2000; Hudson, 1996; McKeown, Beck, Sinatra, & Losterman, 1992).

In summary, current views of the L2 reading process emphasize the importance of linguistic knowledge in L2. It is generally accepted that insufficient knowledge of L2 would constrain L2 processing behaviors and likely result in poor understanding, as it impairs meaning construction based on textual information and restricts higher-level processing such as the use of background knowledge and cognitive strategies (Alderson, 2000; Koda, 2005).

Background knowledge and subject matter/ topic knowledge. In addition to language knowledge, readers' background knowledge (i.e., knowledge that may or may not be relevant to the text content) and subject matter/topical knowledge (i.e., knowledge directly relevant to the text content) affect text understanding and the way new information is recognized and stored (Alderson, 2000). According to schema theory and the interactive notion of reading reviewed earlier, when processing texts, readers' preexisting general and domain-specific knowledge stored in the interlocking mental structures integrates with the new information from the text to fill in the comprehension gap and rapidly construct meaning (Anderson & Pearson, 1988; Rumelhart, 1980). While reading, readers activate their already-existing knowledge automatically and immediately, without which, comprehension would be hindered (Alderson, 2000). Readers' background and topical knowledge play an especially crucial part in L2 academic reading where the reading materials are relatively difficult and the primary concern is to predict the performance on reading tasks involved in academic study (Grabe, 1999; 2002; Hudson, 1996; Urquhart & Weir, 1998).

Cognitive skills. In addition to knowledge, readers need skills to learn and process new information in the text. Cognitive skills have long been held as crucial to reading success. For example, Thorndike (1917) stated that reading was reasoning. He explained that, readers' skills to construct meaning approximated logical inference and deduction, and that good readers thought clearly. Cognitive skills are especially important for L2 readers to solve difficulties in text processing and understanding, such as unfamiliar lexis or complex syntax. Cognitive skills enable readers to use the cues from the text and information in their minds to construct meaning and monitor the process of reading (Alderson, 2000).

Over the last several decades, cognitive skills have been a major area in reading research and various taxonomies of L2 reading skills have been developed (e.g., Carver, 1992a, 1992b; Farhady & Hessamy, 2005; Grabe, 1991; Koda, 1996; Munby, 1978). For example, Grabe (1991) proposed that fluent L2 reading involved automatic word recognition, synthesis and evaluation, and metacognitive skills. These skill taxonomies provide a framework for reading test construction. However, there are several limitations with many of these taxonomies (Alderson, 2000). First, the skills are frequently ill-defined or undefined. Second, the seemingly discrete skills have enormous overlap. For example, in Munby's (1978) skill taxonomy, the skills "identifying the main point or important information in discourse", "distinguishing the main idea from supporting details", and "extracting salient details to summarize an idea" apparently overlap with one another. Third, the existing skill taxonomies lack empirical observation. Despite these criticisms, certain cognitive skills, such as inference, synthesis, and evaluation, have been major components in current models of the L2 reading processes.

Cognitive and metacognitive strategies. In the recent literature on L2 reading and its assessment, cognitive and metacognitive strategies used by ESL learners when processing text and responding to questions have received considerable attention (e.g., Abbott, 2005; Cohen, 1998; Cohen & Upton, 2006; Lumley & Brown, 2004; Phakiti, 2003; Yang, 2000). Further, cognitive and metacognitive strategies have become crucial components in many models of the L2 reading processes (Alderson, 2000; Carrell & Grabe, 2002; Hudson, 1996; Koda, 2005). Research studies frequently found that good readers are more effective in using metacognitive strategies and more capable of describing the use of such strategies (e.g., Block, 1992; Grabe, 1991; Johnston, 1983; Phakiti, 2003). For example, Block (1992) compared the verbal protocols of proficient and less proficient ESL readers in an American college. His finding showed that proficient readers were adept at using metacognitive strategies to answer the reading items, were more aware of how to control the process, and were more capable of verbalizing their awareness than less proficient readers.

In the present study, cognitive strategies refer to the examinees' ongoing mental processes to use their language, background, and topic knowledge to answer the given items, while metacognitive strategies refer to higher order executive processing deliberately used by examinees to direct and control their cognitive processing for successful performance (Phakiti, 2003). Different from cognitive processes that "are likely to be encapsulated within a subject area" such as EFL reading, metacognitive strategies are "thinking about thinking" (Phakiti, 2003, p. 29) and span multiple subject areas (Schraw, 1998). A list of cognitive and metacognitive strategies that L2 readers use during reading include monitoring progress of understanding, planning ahead how to read, and selectively attending to text (Alderson, 2000; Block, 1992; Johnston, 1983).

Purpose and context. In addition to knowledge, skills, and strategies, reader purpose and the context in which L2 readers engage in reading is increasingly being emphasized (e.g., Alderson, 2000; Cohen & Upton, 2006; Enright et al., 2000; Hill & Parry, 1992; Hudson, 1996). These researchers stress that reading is usually undertaken for some purpose and in a specific context, which affects the knowledge and skills required, strategies used, and the understanding and recall of the text. In the context of reading for academic purposes, important reading processes involve "locating discrete pieces of information by skimming and scanning the text", "understanding the main ideas or major points of the text", "constructing an organized representation of the text that includes major points and supporting details", and "integrating information across multiple sources" (Enright et al., 2000, p. 4; Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000).

L2 Reading Test Taking Processes

Test-taking processes refer to the processes used by examinees to complete the tasks on a reading test. When taking a reading test, readers may use the processes that they would not use under non-test conditions (Alderson, 2000; Cohen & Upton, 2006). These may entail the reading processes assessed by the reading tests (e.g., drawing inferences, deducting the meaning of an unknown word, locating specific information in the text) and test management processes (e.g., matching the information in the test item to the information in the text, eliminating the impossible options, and using personal knowledge related to the text topic) (Abbott, 2005; Cohen & Upton, 2006; Drum, Calfee, & Cook, 1981; Farr et al., 1990). In some circumstances, examinees may use short cuts to arrive at answers. For instance, when answering the multiple-choice reading test items, examinees may choose an answer based on their intuition or hunches, their already possessed common sense, the cues in test items, and surface features of the options without reading and comprehending the text. In such cases, examinees may be using testwiseness to circumvent the need to tap their language knowledge and strategic competence that a test is actually measuring (Cohen & Upton, 2006; Gao, 2002; Katz & Lautenschlager, 1995; 2001; Powers & Leung, 1995; Yang, 2000).

The Assessment of L2 Reading

Assessing L2 reading necessarily has some overlap with the nature of the L2 reading processes reviewed in the previous section. In addition, as an assessment issue, assessing L2 reading directly relates to the theoretical framework of language testing and psychometric concerns such as reliability and validity. This section first presents the current theory of communicative language ability (Bachman, 1990; Bachman & Palmer, 1996), which is closely related to the constructs assessed by the MELAB reading (ELI, 2003). Then, the constructs of the L2 academic reading ability are discussed. Finally, the section reviews the literature on the factors affecting the L2 reading performance.

A Theoretical Framework of Communicative Language Competence

In his seminal work *Language and Mind* (1968), Chomsky proposed the notion of *language competence* and distinguished it from *language performance*. According to Chomsky, language competence was limited to linguistic knowledge, while language performance was the actual use of that language in a concrete situation, which was affected by various psychological factors affecting language perception and production. As an alternative to Chomsky's linguistic view of competence, several researchers proposed the notion of *communicative competence*, expanding language competence to the knowledge of the rules of language use in a specific context (e.g., Hymes, 1972; Munby, 1978). Canale and Swain (1980) extended earlier works and proposed the framework of communicative language competence, which included linguistics, sociolinguistic (i.e., the knowledge of how language is used), discourse, and strategic competence (i.e., the ability to use knowledge and competence in a meaningful situation).

Over the last decade, communicative language competence and its assessment have increasingly been emphasized in the language assessment community. Bachman (1990) expanded upon earlier work and proposed the framework of communicative language ability (CLA) for the development and use of language tests and language testing research (see Figure 1).Bachman's framework of CLA acknowledges the language

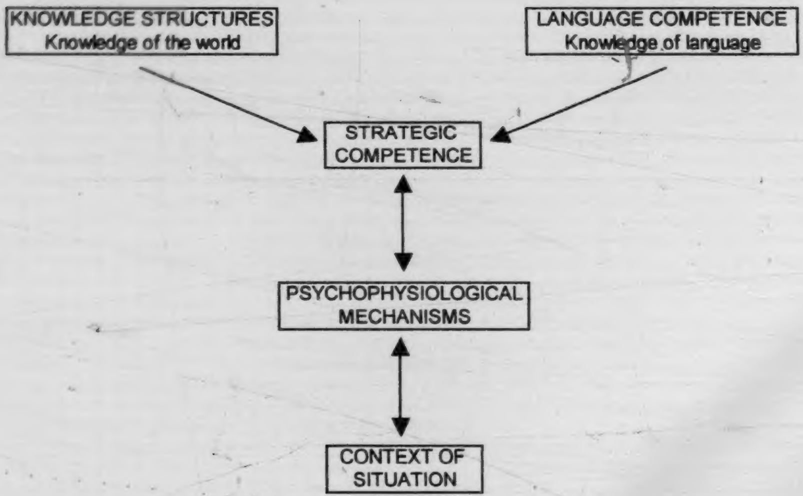

KNOWLEDGE STRUCTURES
Knowledge of the world
LANGUAGE COMPETENCE
Knowledge of language
STRATEGIC COMPETENCE
PSYCHOPHYSIOLOGICAL MECHANISMS
CONTEXT OF SITUATION

Figure 1. Components of communicative language ability (Bachman, 1990, p. 85).

knowledge or competence, the capacity for using this knowledge or competence, and "characterizes the processes by which the various components interact with each other and with the context in which language use occurs" (Bachman, 1990, p. 81). Specifically, this framework consists of language competence, strategies competence, and psychophysiological mechanisms, and describes the interactions of these components with the language user's knowledge structures and language use context. According to Bachman, language competence includes a set of language knowledge components used in communication. Strategic competence performs "assessment, planning, and execution functions in determining the most effective means of achieving a communicative goal" (p. 107), and characterizes the "mental capacity" that relates the language competence to the language user's knowledge structures and the context in which communication occurs (p. 84). Psychophysiological mechanisms are "the neurological and psychological processes involved in the actual execution of language as a physical phenomenon" (p. 84). They characterize "the channel (auditory, visual) and mode (receptive, productive) in which competence is implemented" (p. 108).

Bachman further explains the components of language competence in his framework of CLA (see Figure 2). As Figure 2 shows, language competence includes organizational competence, which consists of grammatical and textual competence, and pragmatic competence, which consists of illocutionary and sociolinguistic competence.

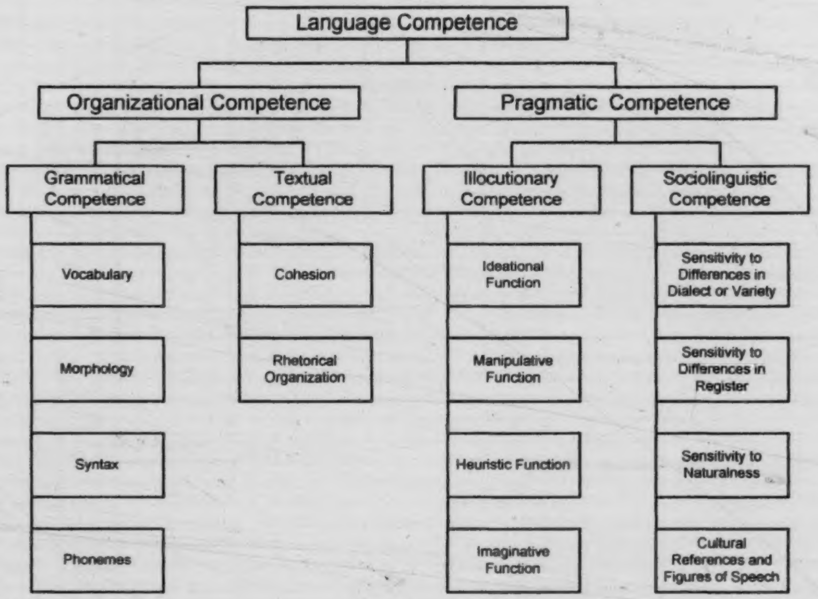

Language Competence
Organizational Competence
Pragmatic Competence
Grammatical Competence
Textual Competence
Illocutionary Competence
Sociolinguistic Competence
Vocabulary
Morphology
Syntax
Phonemes
Cohesion
Rhetorical Organization
Ideational Function
Manipulative Function
Heuristic Function
Imaginative Function
Sensitivity to Differences in Dialect or Variety
Sensitivity to Differences in Register
Sensitivity to Naturalness
Cultural References and Figures of Speech

Figure 2. Components of language competence (Bachman, 1990, p. 87).

Bachman and Palmer (1996) refined Bachman's (1990) framework and clearly defined *language use* as the dynamic creation of intended meanings in discourse by an individual in a particular situation. According to Bachman and Palmer, context and purpose of language use are crucial in defining language ability. As they noted,

> If we are to make inferences about language ability on the basis of performance on language tests…(we) need to define language ability in a way that is appropriate for

> each particular testing situation, that is, for a specific purpose, (a particular) group of test takers, and (a specific) TLU domain (i.e., target language use domain, situation or context in which test takers will be using the language outside of the test itself). (p. 66)

In this study, language use occurs in the context where English language ability of adult non-native English speakers is assessed for academic purposes. This context suggests that the reading ability defined in this study incorporates not only the ability to use language knowledge and strategic competence to solve the test task, but also the ability to apply this knowledge and competence to academic reading in the real world.

In addition to the emphasis of context and purpose, Bachman and Palmer pointed out that language use involves complex interactions among individual characteristics of language users and the interactions among these characteristics and characteristics of language use. Hence, "language ability must be considered within an interactional framework of language use" (p. 61-62). They presented their framework as a theory of the factors that affected performance on language tests and proposed that performance on language tests was affected by (1) the interactions among the examinees' language knowledge, topical knowledge, affective schemata, strategic and metacogntive strategic competence, and personal characteristics such as age and native language, and (2) interactions between examinee characteristics and characteristics of the test task. Subsequently, Bachman (2002) clearly distinguished three sets of factors that affected test performance: examinee attributes, task characteristics, and the interactions between examinee and task characteristics.

The current theoretical framework of CLA (Bachman, 1990; Bachman & Palmer,

1996) is consistent with current views of L2 reading ability and its assessment, which acknowledge the interactive nature of reading and the effect of text and item characteristics, reader knowledge, cognitive and metacognitive strategies, and purpose and context of reading on reading test performance (e.g., Alderson, 2000; Enright et al., 2000; Jamieson et al., 2000).

Assessing L2 Academic Reading Ability

Constructs of the L2 academic reading ability. The current theory of CLA and models of L2 reading suggest a range of constructs of the L2 academic reading ability, which has been operationalised differently in tests of L2 academic reading (Alderson, 2000; Cohen & Upton, 2006; Douglas, 2000; Enright et al., 2000; Hudson, 1996; Jamieson et al., 2000). It has been accepted that word identification skills need to be tested, as word recognition is critical to fluent reading. Knowledge of the language, such as knowledge of vocabulary, sentence structure, and formal discourse structure (e.g., cohesion and rhetorical structure), is essential for L2 readers' understanding of academic texts, and thus should be taken into account in testing L2 academic reading. Cognitive skill and cognitive and metacognitive strategies are important for L2 readers to overcome language difficulties especially when reading difficult academic texts. Hence, L2 academic reading tests should allow examinees to apply their cognitive skills and strategies such as inference, synthesis, evaluation, and monitoring.

Alderson (2000) stresses that in the context of L2 reading, sufficient knowledge of the foreign or second language, cognitive skills, and problem-solving strategies are especially important. Nevertheless, Alderson reminds us that readers' background knowledge is normally not included in the constructs to be assessed, though its influence on L2 reading process and product is recognized.

The MELAB reading. According to the MELAB manual (ELI, 2003), the MELAB reading is designed to measure examinees' understanding of reading texts in college-level academic settings. The constructs assessed by the MELAB reading test items reflect current understanding of the L2 academic reading ability and its assessment reviewed above. Specifically, the constructs assessed by the MELAB reading include recognizing the main idea and supporting details of the text, understanding the relationship between sentences and portions of the text, recognizing the cohesive devices, organizational pattern, and argument method of the text, drawing text-based inferences, understanding pragmatic and rhetorical purposes of the text creator, and recognizing specific vocabulary in context. The reading section consists of four passages, with each followed by five multiple-choice items, for a total of 20 items. Each item consists of a question stem and four options (1 key and 3 distracters). Examinees are instructed to read the passages and then select the single best answer based on the information in the given passage.

In addition to the constructs, text and item characteristics are important considerations in assessing academic reading ability (Cohen & Upton, 2006). According to the specifications for the MELAB reading, all test passages are expository texts adapted from publications of general interest to educated adults and edited to make them coherent, clear, self-contained, and free of jargon. The total length of the passages included in each test form is 922-1021 words, with each passage containing approximately 250-300 words. The style of the language characterizes English for academic purposes. The readability of passages, as measured by a standard readability formula, suggests that the vocabulary size and structural complexity of the passages are at the college level. The genres of the passages include humanities (literature, folktales), social science (anthropology, history, government), physical science (astronomy, physics, mechanics), and biological science (biology, zoology, medicine). The topics of passages are accessible to all examinees; no specific prior knowledge is required to understand a passage or answer an item. Unfortunately, controlling for prior knowledge in assessing reading has been difficult, if not impossible, in most studies of reading tests (e.g., Carr, 2003). Research has shown that for examinees having reached a certain minimum reading proficiency, prior knowledge possessed in the content area of a reading text can have a facilitating effect while deficit in text-related prior knowledge may increase the demands on processing, understanding, and recall of that text (Alderson, 2000; Clapham, 1996; Grabe, 2004; Phillips, 1988). In an attempt to counter any possible bias towards examinees of a particular educational or cultural background, ELI selects texts on a range of topics and includes different genres of passages in each test form.

According to ELI item-writing guidelines, the questions following each passage are intended to assess a variety of academic reading abilities assessed by the MELAB reading. The set of questions cover the entire passage rather than a portion of it, and assess understanding of the important information rather than insignificant details. The questions recognize the complexity of the passage, but the questions themselves are short, clear, and simple in vocabulary and syntax so that they are easy to read and understand. It is emphasized that questions should be answered based on relevant information in the passages rather than by analysis of the question structure, information disclosed from response options, or examinees' knowledge outside the passage.

Variables that Affect the Difficulty of L2 Reading Test Items

Prior research has identified several potential sources of processing difficulty that may affect the difficulty of L2 reading test items and suggested a number of variables to score the cognitive complexity of these sources (e.g., Bachman, Davidson, Ryan, & Choi, 1995; Bachman, Davidson, & Milanovic, 1996; Enright et al., 2000; Jameison et al., 2000; Perkins & Brutten, 1992; Rupp, Garcia, & Jamieson, 2001). Consistent with the emphasis of the reader-text interaction and the importance of test method in current literature of the L2 reading and its assessment (e.g., Alderson, 2000; Bachman & Palmer, 1996), the variables that affect the difficulty of L2 reading test items can be classified into two categories: task (text and item) variables and reader variables. Examples of task variables include text topic, text type (e.g., expository, narrative) and genre (e.g., magazine article, feature story), text organization, linguistic features (e.g., syntactic complexity, vocabulary difficulty), text readability and length, sufficiency and clarity of the given information, occurrence of distracting information in the text, language of questions, frequency and usage of particular words involved in the task, and concreteness of the requested information (Alderson, 2000; Bachman et al., 1996; Carr, 2003; Enright et al., 2000;

Freedle & Kostin, 1993; Jamieson et al., 2000; Skehan & Foster, 2001). Examples of reader variables include readers' language knowledge (e.g., vocabulary and syntactic knowledge), background and topic knowledge, cognitive skills (e.g., drawing inferences), problem-solving strategies (e.g., monitoring), reader purpose (e.g., for entertainment or test taking), reader affect (e.g., motivation, interest), and other reader characteristics (e.g., age, personality) (Alderson, 2000; Enright et al., 2000; Jamieson et al., 2000; Skehan & Foster, 2001).

Alderson (2000) commented that task variables and reader variables that affect the L2 reading performance are like two sides of a coin in the reader-text interaction. The task variables simultaneously interact with the reader variables, which affects the process of reading and the difficulty of reading test items. For example, Clapham (1996) investigated the effect of text topic and content on L2 reading comprehension, and found that the topic of a passage interacted with both readers' background knowledge and their L2 language knowledge.

In terms of the multiple-choice (mc) reading test items which has specific relevance to the current study, the variables that can affect item difficulty have taken into account task characteristics, which involve characteristics of text, question stem, correct option, and distractors, and attributes of test-takers, which involve how they might process the task and how familiar they might be with the task content (Alderson, 2000; Bachman, 2002; Kasai, 1997). Text characteristics are considered, because the difficulty of a reading test item has been considered a function of item characteristics and the text on which the item is based (Bachman et al., 1996; Enright et al., 2000). Example task variables that affect the processing difficulty of mc test items include availability of context cues, vocabulary difficulty, complexity of syntax and text structure, type and location of requested information, text topic, degree to which the correct answer and the question stem match the wording of the information in the text, and plausibility of distractors. Examples of reader variables that affect the cognitive complexity and hence the difficulty of mc items include readers' knowledge of vocabulary, syntax, text structure, and their inference and synthesis skills (Alderson, 2000; Bachman et al., 1995; Enright et al., 2000; Huff, 2003; Jameison et al, 2000; Kasai, 1997; Sheehan & Ginther, 2001). The variables scored for consideration in the current study were based on both theoretical relationships suggested by the literature reviewed above and empirical relationships revealed by the studies on the factors that affect the L2 reading test item difficulty, which are reviewed in the next section.

Research into the Factors Affecting L2 Reading Test Item Difficulty

Understanding the factors that affect task performance, especially the construct-relevant sources of task difficulty, is the major goal of a theory in a particular domain (Messick, 1995) and the most pressing issue for language testing (Bachman, 2002). This issue is central because it guides the development and scoring of test tasks, and makes possible theory-based score interpretations and test use (Bachman, 2002; Embretson, 1998; Skehan, 1998). This section discusses the methods and issues concerning the study of the factors affecting the L2 reading test item performance, and critically reviews the representative studies. This section is intended to inform the concepts and methods for the current study and lead to a discussion of the psychometric literature of specific relevance to modeling the cognitive processes underlying reading test item performance.

Methods and Issues Concerning This Research

Over the last decade, language testers have been studying the factors that affect reading task or item performance, especially the factors affecting the difficulty of multiple-choice reading test items. While a considerable amount of research has been devoted to the L1 context (e.g., Drum et al., 1981; Farr et al., 1990; Embretson & Wetzel, 1987; Gorin, 2002; Kirsch & Mosenthal, 1990; Kintsch, 1998; Sheehan, 1997; Singer & Kintsch, 2001; VanderVeen, 2004), only a few studies have been conducted in the L2 context (e.g., Bachman et al., 1996; Carr, 2003; Freedle & Kostin, 1993, 1999), and many of these studies are limited in terms of the concepts and methods used.

Research into the factors affecting the L2 reading test item performance has generally referred to three sets of factors: (1) task characteristics that are considered to be "inherent in the task itself" and independent of readers, (2) reader attributes, such as how readers process the task and how familiar they are with the text topic, and (3) interactions between readers and task characteristics (Bachman, 2002, p. 469). A variety of empirical methodologies have been used, ranging from the qualitative analyses of item content and item-solving processes to the more commonly used quantitative analyses, such as multiple linear regression and factor analysis. In addition, a variety of reading tasks have been used that vary in topic, genre, vocabulary difficulty, syntactic complexity, textual organization, and purpose. This research has yielded a number of factors that appear to affect item performance across a variety of reading tests. However, limited by the concepts and methods employed, the research results, as a whole, have been equivocal and inconsistent (Bachman, 2002).

Conceptually, as reviewed earlier, current theories of reading recognize the interactions between reader and text and emphasize the purpose and context of reading (Alderson, 2000). Moreover, current theories of language testing consider task performance as a function of interactions between examinee attributes and test task characteristics (Bachman, 2002). However, many of the existing studies of the factors that affect the L2 reading test item performance either focused on the characteristics inherent in text and/or the item itself without taking examinees into account, or vice versa. In addition, varying purposes and contexts of reading tasks were not given proper attention in many of these studies. For example, a factor affecting the difficulty of reading test items in non-academic settings may not function similarly in academic settings. Methodologically, many of these studies are limited in the analyses employed. Consequently, the results of this research do not seem to advance our understanding of the L2 reading test item performance (Bachman, 2002). Empirical studies of the factors affecting the L2 reading test item performance are reviewed below along four lines, according to the concepts and methods used.

Studies of Surface Task Characteristics and Item Performance

Studies within this group typically identify a number of text and/or item characteristics that are considered to be independent of readers, and then investigate the effect of these characteristics on item statistics using quantitative methods, such as the commonly-used multiple linear regression (e.g., Freedle & Kostin, 1993, 1999). These studies have suggested a number of text and item characteristics that may affect item performance across a variety of L2 reading tests, which have clear implications for the design of L2 reading tests. However, due to over-reliance on surface characteristics of texts and items without taking into account reader factors in item solving, the analyses provide neither detailed descriptions about the processes used by examinees during test-taking, nor meaningful feedback on what a reader should know and do to correctly answer a given item (Alderson, 2005b). In addition, the multiple linear regression analysis method used in many of these studies has its limitations, which may affect the accuracy of results and may not be satisfied in practice, such as oversensitivity to the presence or absence of an item characteristic variable (Kasai, 1997) and strict requirements for linearity and the number of items (Keppel & Zedeck, 2001).

Freedle and Kostin (1993) examined the effect of task characteristics on the difficulty of TOEFL reading items, as measured by equated delta ($n_{items} = 213$; $n_{examinees} = 2000$). Based on a review of previous studies predicting the difficulty of multiple-choice reading test items, they hypothesized that 12 categories of 65 text, item, and text-by-item interaction variables might influence the difficulty of TOEFL reading items. After a multiple-regression analysis, they found that 58% of the variance in item difficulty was explained by eight categories of text and text-by-item variables: negations, referentials, rhetorical organizers, sentence length, paragraph length, passage length, lexical overlap between text and options, and location of relevant text information. Their investigation of reading item difficulty as a function of text, item, and text-item interaction impacted later research and their findings have direct implications for text writing and item design for L2 reading tests. However, the variables used in their study, which were mainly word counts (e.g., the number of words in correct option, the number of referentials summed over incorrect options), fail to reveal examinees' complex processes of item solving and lack interpretive and diagnostic value (Kasai, 1997).

Carr (2003) examined task characteristics in explaining the difficulty of 146 reading items included in three TOEFL test forms. Based on a review of previous research, he developed a rating instrument consisting of three sets of 311 passage, key sentence, and item variables (e.g., topic, rhetorical features, and cohesion of the passage, length and location of the sentences in the passage that contain key information required to correctly answer a given item, and characteristics of question stems, options, and the interaction between item and passage). He asked five graduate students in applied linguistics to rate the task characteristics using the rating instrument. Similar to Freedle and Kostin's (1993) study, most of the variables used in Carr's investigation were word and sentence counts. However, Carr considered text characteristics of a reading test as most relevant to fluent reading and most reflective of the target language use domain (Bachman & Palmer, 1996), but considered item characteristics as less authentic aspects of the test. Hence, only text (passage and key sentence) variables were included in his analyses. Through exploratory and confirmatory factor analyses, he constructed and tested a factor model of the text characteristics and concluded that passage content, syntactic features of key sentences, and vocabulary factors contributed to the difficulty of the TOEFL reading items.

Carr provides a thorough list of text variables that may affect the difficulty of L2 reading test items and an alternative method for investigating the effect of text characteristics on reading item difficulty. However, excluding item variables from the analysis does not seem to be warranted, since the complete task of multiple-choice reading tests involves text, question stem, and options, and examinees' mental processes used to answer multiple-choice reading test items may differ from those used to answer constructed response or essay questions (Kasai, 1997). Hence, item characteristics (question stems and options) need to be considered in the case of multiple-choice tests. In addition, like Freedle and Kostin's (1993) study, a focus on the surface characteristics of text fails to describe in detail examinees' cognitive processes during test-taking. Finally, the rating instrument consisting of 311 variables greatly increases the workload of the raters. In practice, a complicated rating instrument as such may not be feasible.

Studies of Cognitive Demands of Test Items and Item Performance

Studies within this group typically identify 'item features' that are essentially cognitive demands hypothesized to affect the performance of a given item (e.g., Alderson, 1990a; Alderson & Lukmani, 1989; Bachman et al., 1995; Bachman et al., 1996; Skehan, 1996, 1998). These studies used 'expert' ratings of the test items that included different combinations of the cognitive demands, and then related the ratings to item performance using cross-table or multiple linear regression analysis. 'Experts' employed in these studies have included various individuals, such as EFL teachers or administrators and graduate students in applied linguistics or educational psychology. The results of these studies consistently indicate no systematic relationship between 'expert' ratings and item statistics. The equivocal results are likely caused by methodological limitations. For example, none of these studies examine the relationship between 'expert' ratings of the cognitive demands of a given test item and item statistics using advanced measurement models that incorporate the cognitive elements of items. In addition, item statistics calculated using the classical test theory model have little connection with the cognitive processes used to answer an item (Embretson, 1999). Finally, 'experts' may process the test tasks differently from the target examinees (Alderson, 2000, 2005a; Leighton & Gierl, 2005). Despite the limitations, these studies begin to pay attention to the effect of cognitive elements of test items on item performance, which anticipates the cognitive processes used by examinees when they answer test items and preludes the study of item performance in light of examinees' actual cognitive processes. In addition, 'expert' analysis may reveal both automatic and controlled processes evoked by test items (Leighton, 2004). As automatic processes are inaccessible for description through conscious verbal reports (Cheng, 2003), the analysis of the cognitive demands of a test item provides valuable sources of data to supplement verbal reports.

Alderson and Lukmani (1989) investigated the cognitive skills required for correctly answering the reading items included in a L2 communication skills test taken by 100 students at Bombay University (India), and related the skill requirements of individual items to item difficulty, as measured by percentage of correct responses. Nine teachers at Lancaster University (Great Britain) were asked to describe what was being tested by each of the 41 test items. Results showed little agreement among the judges on the skills being tested by each item and little relationship between item difficulty and the skill requirements of each item. Likely reasons for these equivalent results may include the lack of a pre-structured rating guide and pre-training of the judges. In addition, correctly answering an item may involve multiple skills. Moreover, the judges at Lancaster may not have been familiar with how students at Bombay processed the test task.

Using a rating instrument containing 14 reading skills, Alderson (1990a) conducted a similar study, in which 18 teachers of ESL were asked to decide the single skill being tested by each of the 15 short answer questions on two British language proficiency tests. Again, little agreement was reached among the judges and little relationship was found between item difficulty and skill requirements of the items. Two likely reasons for the equivocal results are: (1) correctly answering an item may require multiple skills, while the judges were allowed to specify only one skill for each item, and (2) the enormous overlap among the skills provided on the rating instrument may have affected the accuracy of expert rating.

The two studies reviewed above question the ability of expert judges to determine the skills being tested by a given item. However, other studies have reported high levels of agreement among expert judges when using well-designed and clearly-defined rating instruments, extensive discussion, exemplification, and re-categorization of the skills (e.g., Bachman et al., 1995; Bachman et al., 1996; Carr, 2003; Lumley & McNamara, 1995; Xi, 2003). In Bachman et al.'s (1996) study, five trained applied linguists with experience as EFL teachers were asked to analyze the characteristics of 25 vocabulary and 15 reading items and passages on each of the six parallel forms of an EFL test from the University of Cambridge Local Examinations Syndicate. The number of examinees for each form ranged from 431 to 1099. A refined rating instrument was presented to the raters, which contained 23 test task characteristics (TTC) and 13 components of communicative language ability (CLA) defined using Bachman's (1990) framework. The rating scales were ordered in the way that higher ratings corresponded to easier items, with TTCs scaled from 0 to 2 and CLAs scaled from 0 to 4 (see Bachman et al., 1995 for the full rating instrument). Rater agreement was checked using generalizability analysis and the rater agreement proportion, a statistic of agreement. The results showed that the overall rater agreement was very high and that the TTC ratings were more consistent than the CLA ratings. Content comparability and statistical equivalence of the forms were checked using descriptive statistics. Results supported the comparability of the forms. Finally, they related the TTC and CLA ratings to the IRT item parameter estimates calibrated using the 2PL model and the PC-BILOG version 3.04 computer program (Mislevy & Bock, 1990). Step-wise regressions were performed for all items and for vocabulary and reading items separately, by individual form, and for all forms combined. Results showed that neither TTC nor CLA ratings consistently predicted item difficulty or discrimination across the six test forms, though a combination of the TTC and CLA ratings consistently yielded high predictions.

Bachman et al.'s study demonstrates the possibility of achieving high-level agreement among judges and provides some evidence for the relationship between item characteristics and item statistics. It appears that their use of a rating instrument and rater training plays an important part in rater agreement. Several limitations with their study warrant improvement. First, more refined definitions of the abilities may increase the consistency of ability ratings. Second, the inconsistent prediction of item parameter estimates across the forms indicates that identifying item characteristics is likely affected by differences among tests in the passages used and the nature of items included. If that is the case, then a large number of tests may be examined to provide reliable item characteristics that affect item performance (Alderson, 2000). Finally, as 'experts' may process the test tasks differently from the target examinees, it is imperative to investigate examinees' actual processes underlying the correct responses (Alderson, 2005a; Leighton & Gierl, 2005). As Alderson (2000) reminds us, "what matters is not what test constructors (or experts) believe an item to be testing, but which responses are considered correct, and what process underlies them" (p.97).

Processes in Task Performance Inferred from Verbal Reports

With the development of cognitive psychology, introspection, particularly in the form of concurrent verbal reports (i.e., an individual's description of the processes he/she is using during task solving) and retrospective verbal reports (i.e., the recollection of how the task was solved), has been established as a valid means to obtain valuable sources of data on cognitive processing during task performance (Ericsson & Simon, 1993). Recent calls for the integration of cognitive psychology and test theory have resulted in a revival of the verbal report method for investigating the processes in task performance (Leighton, 2004). Leighton (2004) recommends the collection of both concurrent and retrospective reports to triangulate the processes actually used by examinees to think about and solve the problems. She further identified the conditions critical to the successful use of verbal report, such as using tasks of moderate difficulty to maximize the verbalization elicited, conducting analyses of a task's cognitive demands prior to eliciting verbal reports to anticipate the cognitive processes a respondent will use when solving the task. The last decade has seen an increasing use of verbal reports to inspect the processes of L2 readers during test taking (e.g., Abbott, 2005; Alderson, 1990b; Allan, 1992; Anderson, 1991; Anderson, Bachman, Perkins, & Cohen, 1991; Block, 1986; 1992; Cohen & Upton, 2006; Lumley & Brown, 2004; Phakiti, 2003; Yang, 2000). These studies have shed some light on the cognitive processes underlying L2 reading test item performance and suggested a number of cognitive processes that appear to predict item statistics on a variety of L2 reading tests. However, as the test tasks and statistical analyses employed differ widely across the studies, these studies have produced different results on the cognitive processes that affect item statistics on the L2 reading tests.

Alderson (1990b) conducted a pilot study to inspect the cognitive processes used by examinees when they answered 10 short-answer questions in a L2 reading test. The participants were two graduate students studying English for academic purposes in the intermediate class at the Institute for English Education of the University of Lancaster. Concurrent and retrospective verbal reports were collected, with one student taking the test on his own and then being interviewed how he had answered the questions and the other thinking-aloud while answering the questions. Despite the limitations in scope and the way in which the verbal reports were collected (e.g., the lack of probing questions to systematically explore the processes), Alderson's study reveals several interesting findings. First, the processes used to answer an item involve multiple skills. Second, correctly answering an item was often associated with test methods and knowledge of particular lexical items. Third, the relationship between cognitive processes and item difficulty is "far from clear, but is certainly not simple" (p. 478).

Anderson et al. (1991) investigated the strategies used by adult EFL learners to complete two forms of a standardized reading test, and then examined the relationships among test-taking strategies, item content, and item performance using a triangulation of three sources of data: retrospective verbal reports, item content, and item performance. The test consisted of 45 multiple-choice test items based on 15 reading passages. The passages were 44 to 135 words in length, on a range of topics, and in a variety of styles. The questions were designed to measure three types of reading skills: recognizing main ideas, understanding direct statements, and drawing inferences. Both forms of the test were administered to a group of 28 Spanish-speaking students enrolled at a university-level ESL program. The students were classified as beginning (9), intermediate

(10), and advanced level (9) according to the placement test by the program. They ranged in age from 18 to 34 years and had studied in the US from nine weeks to nine months. The students were randomly assigned to two groups, with one group taking Form A and the other taking Form B. After the test, the students were introduced to the verbal report procedures and allowed to practice. One month later, the alternate form of the test was given to the same group of students. The students were told to (1) read the passages and answer the questions associated with each passage, and (2) retrospectively verbalize, in their L1 (Spanish), L2 (English), or both, the strategies they used while reading the passages and answering the questions.

The verbal reports data were transcribed and coded for strategies in five categories: supervising, supporting, paraphrasing, establishing coherence, and test-taking strategies. Each processing strategy was defined and examples were provided. To investigate the reliability of assigning strategies to the categories, two raters independently classified data for ten randomly selected verbal reports. Their classifications were compared to those of the researcher, and the percentage of agreement across all three raters was 74% (i.e., the number of times that the raters agreed on the categorization of each reported strategy). Next, they conducted a content analysis of the items on the two test forms, based on test specifications (main idea, direct statement, and inference) and Pearson and Johnston's (1978) taxonomy of text-item relationships (textually explicit, textually implicit, and scriptally implicit). Then, they examined test performance data. The items from the two administrations of the test were scored and the item difficulty (*p*) and discrimination (r_{pbi}) were calculated. Finally, chi-square analyses were conducted to examine (1) the relationship between strategy use and item type, (2) the relationship between item type and item difficulty, and (3) the relationship of strategy use to item difficulty and discrimination.

Their results revealed a significant relationship between frequencies of the reported strategies and item type. For example, the strategy "match stem with text" was reported more frequently for inference items than for main idea items. Their results also indicated a significant relationship between frequency of strategy use and item difficulty. The strategies for which the significant relationships occurred included skimming, guessing, paraphrasing, responding affectively to text, matching stem with text, selecting answer through elimination, selecting answer because stated in text, selecting answer based on understanding text, and making reference to time. In addition, their results showed that more strategies were reported for the items of average difficulty ($0.33 \leq p \leq 0.67$) than for the difficult items ($p < 0.33$), and that more strategies were reported for the difficult items than for the easy items ($p > 0.67$). This finding appears to support the use of moderately difficult items to maximize the verbal report data (Leighton, 2004). However, no significant relationship was discerned between the item type and item difficulty. Anderson et al.'s study demonstrates the use of a triangulation approach to the construct validation of a standardized reading test: the test developers' analysis of the item content clarifies the constructs assessed by the test; the data inferred from verbal reports provide additional insights into the cognitive processes of L2 readers during test taking; the performance data provide a better understanding of the test. The authors recommend the use of multiple data sources and stress supplementing the traditional psychometric approach with qualitative analysis of item content and verbal reports. Their study offers considerable promise for further research on large-scale standardized reading tests.

Phakiti (2003) investigated cognitive and metacognitive strategy use in relation to the EFL reading test performance, using a combination of quantitative and qualitative analyses of three sources of data: questionnaire, test performance, and retrospective interview. In his study, 384 students enrolled in a fundamental EFL course at a Thai University took an 85-item multiple-choice reading comprehension achievement test in a 3-hour period and then answered a 35-item cognitive-metacognitive questionnaire on the strategies they used while completing the test. The questionnaire used a 5-point Likert scale and included strategies such as "I used my own English structure knowledge to comprehend the text", "I tried to find topics and main ideas by scanning and skimming", and "I asked myself how the test questions and the given texts related to what I already knew" (pp. 55-56). According to the test performance and teacher judgment, the participants were classified as highly successful, moderately successful, and unsuccessful, from which four highly successful and four unsuccessful students were selected for retrospective interviews. They were asked to report retrospectively the strategies used when they completed the test. Then, they took a 10-minute multiple-choice reading test with one passage and six items and described about the strategies they used retrospectively. All interviews were conducted in Thai and lasted about 30 minutes.

The findings of Phakiti's study revealed three major points. First, cognitive and metacognitive strategies explained, respectively, 15% and 22% of the test score variance, and had a positive relationship to the reading test performance ($r = 0.391$ and 0.469). Second, highly successful readers reported the use of metacognitive strategies more frequently than the moderately successful readers, who in turn reported the use of metacognitive strategies more frequently than the unsuccessful readers. Third, test-taking strategies were more frequently reported than reading strategies. However, his study was limited to the strategies, which are evoked only when examinees are faced with problems or difficulties that cannot be resolved by their automatic skills (Phakiti, 2003). Factors other than strategies that may result in different processing of examinees and affect task performance, such as text characteristics and knowledge and skill demands of the test items, are not taken into account.

Another study of L2 readers' strategy use and item performance using a combination of qualitative and quantitative methodologies is Abbott's (2005) investigation of differential item functioning with the Canadian Language Benchmarks Assessment (CLBA) across two cultural groups, Chinese and Arabian. In order to identify the strategies involved in answering each of the 32 CLBA reading items, non-mediated concurrent and retrospective verbal reports were collected from eight intermediate ESL learners of each cultural group. Due to the concern that 32 items was too long for verbal reporting, data were collected on two sessions during the same week. In each session, the participant was introduced to the verbal report procedures and provided with an opportunity to practice. Then, the participant was asked to think-aloud while working through the options and deciding the answer. Upon completing each item, the participant was asked to explain what he/she had thought and done to arrive at the answer. Supplementing the verbal reports, a background questionnaire and a questionnaire on topic familiarity and understanding of each item and passage were administered before and after the verbal report, respectively.

The audio-taped data were transcribed and translated into English, segmented and coded into seven categories of bottom-up and top-down strategies. Consistency of the coding was examined by having an independent rater code about 35% of the total items. The degree of agreement between the researcher and the rater was 90.6%. Next, three ESL reading experts independently classified each of the 32 reading items into one of seven strategy categories that were identified from the verbal report analysis. The items were coded based on expert judgment of the strategy most critical to correctly answering the items. The coding results of the three raters were compared; inconsistencies were discussed and consensus reached. Finally, the items were grouped into bundles based on the consensus codes to conduct the differential bundle function analyses.

Her findings revealed significant group differences in four bottom-up strategy categories and three top-down categories. The bottom-up strategies, "breaking words into smaller parts", "scanning for details", "identifying synonyms or paraphrases", "and matching key vocabulary in the text to key vocabulary in the item" were found to favor the Chinese examinees, while the top-down strategies, "skimming for gist", "connecting or relating information presented in different parts of the text", and "drawing inferences based on information presented in the text", were found to favor the Arabian examinees.

Abbott's study has a couple of limitations. First, in her study, experts found it hard to classify the items into a specific category. A likely reason is that each item was coded for a single strategy, while correctly answering a reading test item often requires multiple cognitive processes (Alderson, 1990b; Huff, 2003; Kasai, 1997). Second, a coding scheme based on the framework that has a clear-cut distinction between bottom-up and top-down processing does not seem to be consistent with current theories of reading, which emphasize the consistent and simultaneous interaction between bottom-up and top-down processing (e.g., Alderson, 2000; Stanovich, 2000). Hence, her coding scheme failed to provide explicit and accurate definitions of the cognitive demands of the items, and as a result, it was frequently found that both bottom-up and top-down strategies were critical to correctly answer the items. A fine-tuned coding scheme representing more complex cognitive demands of the reading items and current theories about the constructs of L2 reading may enhance the interpretability of the results. Despite the limitations, Abbott's study again demonstrates the value of using multiple sources of data and combining both qualitative and quantitative analyses. More significant, it takes the initiative in integrating cognitive psychology, L2 reading, and measurement, which has direct implications for the item difficulty research.

The last study reviewed in this group is Cohen and Upton's (2006) study of the reading and test-taking strategies that examinees used to complete the *Next Generation*

TOEFL reading tasks, using the verbal report method. To refine the coding scheme and procedures, they conducted a pilot study, which consisted of four subjects enrolled in the pre-academic ESL courses. The pilot results showed that the reading tasks were well beyond the subjects' reading ability and the strategies they reported were mainly testwiseness strategies such as random guessing and using item design clues and assumptions. Hence, in their main study, only the students who scored above 20 out of 42 points on the reading test were included. Concurrent and retrospective verbal reports were collected from 32 non-native speakers of English attending the undergraduate or graduate program at the University of Minnesota. These students represented four language groups (Chinese, Japanese, Korean, and Other) and had stayed in the US for four to sixty months. The verbal reports were coded using the coding scheme for reading, test management, and testwiseness strategies developed from the literature and refined through the data coding process. The frequency of strategy use was summarized across all item types and the item-solving processes were analyzed across the respondents. Their study revealed several findings. First, the respondents approached the reading tasks as a test-taking task rather than a reading task. That is, they focused on solving the items instead of learning information from the text. Second, the respondents tended to answer the questions based on their reading and understanding of the passages rather than on their background knowledge. Third, the reported strategies were generally consistent with the constructs of academic reading abilities assessed by the test. Lastly, the differences of strategy use were primarily due to proficiency level, not to language group. Overall, they concluded that the test was testing what it was purported to test. Their study reveals the processes used by examinees in responding to the *Next Generation TOEFL* reading tasks and sheds some light on the construct validation of this test. However, their study failed to include other cognitive processes in use that were not described in the verbal reports. A second source of data on the processes of item solving, such as a strategy questionnaire or raters' analysis of the processes required to correctly answer a given item, may reveal more processes that supplement the verbal report data.

Item Modeling with New Concepts and Methods

Due to conceptual and methodological limitations discussed earlier, current approaches to understanding test item performance are "unlikely to yield consistent or meaningful results" (Bachman, 2002, p. 468). To overcome these limitations, new concepts and measurement models are required. A recent call for the union of cognitive psychology and measurement has seen a revived interest in the effect of a task or item's cognitive demands on task or item performance within the language testing community (e.g., Brindley & Slatyer, 2002; Norris, Brown, Hudson, & Bonk, 2002; Skehan & Foster, 2001). In addition, there has been a growing psychometric literature on modeling test item performance in light of the cognitive elements of an item (e.g., Embretson & Wetzel, 1987; Gorin, 2002; Huff, 2003; Rupp et al., 2001; Sheehan 1997; Sheehan & Ginther, 2001). Studies in this vein typically rely on expert analysis of cognitively-based item features (i.e., item characteristics associated with the cognitive processes involved in item solving), and then relate these features to item statistics using new measurement models that can incorporate such features. These studies have demonstrated that modeling item performance in light of the cognitive processes underlying item solving has many advantages such as informing test design and validation and providing detailed diagnostic feedback to benefit instruction and learning. As Wainer et al. (2000) suggested, linking "indicator variables that distinguish the cognitive processes assumed to be involved in item solving" and "observable item performance indices, in particular, item difficulties" can provide invaluable validity information and rich sources of data for understanding the cognitive processing during task performance (p. 114).

However, there are several limitations with some of these studies. First, the cognitively-based item features are simply judged by 'experts', without being validated by examinees' rendition of the actual processes they used while answering the items. As mentioned earlier, as testers or raters may process the task in a way different from the target examinees, 'expert' judgment about the cognitive processes required to answer an item may not represent examinees' actual processes underlying item performance. Second, item parameter estimates calibrated using the 2-PL or 3-PL IRT measurement models are problematic in the case of passage-based testlets. This is because the interrelatedness among the set of items based on a common passage violates the local item independence assumption of IRT, which can cause inaccurate estimation of examinee abilities and item parameters (Kolen & Brennan, 2004; Lee, 2004; Wainer & Lukhele, 1997). However, no studies have attended to these problems. Third, due to the gap between cognitive psychology, measurement, and reading, many of these studies fail to incorporate the most current cognitive theories in reading or justify the item features within a framework of ability constructs, which are critical for defining item features and interpreting the models. Further, due to technical complexity, only a few studies have explained reading test item performance using new measurement models, and these studies have mainly been devoted to L1 reading. As Cohen and Kolstad (2000) observed, current assessment practice either applies 20th century measurement models to 19th century substantive theories, or vice versa. More research into the L2 reading test item performance through more integration of current L2 reading theories with new measurement models is sorely needed. Recent psychometric literature modeling reading item performance with cognitively-based item features and the tree-based regression measurement model does offer considerable promise for understanding the L2 reading test item performance.

The Tree-based Regression (TBR)

This section discusses how the TBR measurement model can be used to model cognitive processing underlying reading item performance. The section begins with an introduction of the TBR, in which the definition, purpose, and procedures of the TBR are presented and advantages and issues associated with the TBR are addressed. Then, it critically reviews illustrative studies modeling item performance using the TBR.

A Description of the TBR

TBR is a nonparametric technique for classifying cases into homogenous groups. In educational measurement, TBR has been successfully used to model the nonlinear ways in which cognitive demands of a test item interact with various features of the item to predict item statistical properties such as item difficulty, discrimination, and guessing (e.g., Enright, Morley, & Sheehan, 2002; Huff, 2003; Sheehan & Ginther, 2001). These studies have demonstrated that by estimating the probability that examinees at specified score levels will respond correctly to items requiring specified combinations of cognitive structures, the item difficulty model developed with the TBR can be used to draw inferences about examinees' proficiency and provide empirical evidence for the effect of combined cognitive structures on task performance (Sheehan, 1997; Wainer et al., 2000). As the TBR involves specific patterns of skill mastery underlying examinees' observed item responses, it can provide descriptive diagnostic information and useful feedback for instruction and learning (Sheehan, 1997).

Similar to classical regression, TBR provides a method for predicting the value of the criterion, *Y*, from a set of classifications or predictors, *X*. In the case where TBR is used for item modeling, *Y* is the vector of item difficulty or discrimination estimates, and *X* is the matrix of hypothesized skill classifications for each item. The elements of *X* can be expressed on a binary scale (e.g., $x_{ij} = 1$ if skill *j* is required to correctly answer item *i*, and 0 otherwise), on a multilevel categorical scale (e.g., $x_{ij} = A$ if item *i* belongs to schema *A*, $x_{ij} = B$ if item *i* belong to schema *B*), or on a continuous scale (i.e., continuous numeric measures, such as vocabulary difficulty).

Compared to classical regression, TBR has three unique features. First, unlike classical regression where the criterion is predicted by a linear combination of the predictors, the TBR predicts the values of *Y* by clusters of observations sharing similar values of X, and is particularly useful when non-linear relations and higher-order interactions are expected (Introduction to AnswerTree, 2002). Second, unlike classical regression, the TBR does not need to specify interaction terms and is less affected by outliers. Third, unlike classical regression where generating a predicted value for Y is the main purpose, the TBR uses the prediction rule to classify observations into homogeneous sets so that the resulting model is easier to interpret.

In TBR analysis, clusters are identified by successively splitting the observations into increasingly homogeneous subsets called nodes. A recursive partitioning algorithm (Breiman, Friedman, Olshen, & Stone, 1984) is used to determine the optimal variable on which the observations are split into two nodes. The algorithm can be used for either user-specified splits or data-optimized splits. At each stage of the analysis, the original set of observations is called the parent node and the two subsets are referred to as the left and right child nodes.

To illustrate the TBR analysis, consider an item by skill matrix, X, containing a set of 28 items and a single binary-scaled skill classification. The prediction rule for this input is:

$$\text{If } x_i = 0, \text{ then } \hat{y}_i = \bar{y}_0, \tag{1}$$

$$\text{If } x_i = 1, \text{ then } \hat{y}_i = \bar{y}_1, \tag{2}$$

where $\bar{y}_0$ is the mean of y based on all items coded as not requiring skill x (i.e., $x_i = 0$), and $\bar{y}_1$ is the mean of y based on all items coded as requiring skill x (i.e., $x_i = 1$). According to this prediction rule, items coded as requiring skill x are classified into one cluster, and items coded as not requiring skill x are classified into the other cluster. Hence, suppose 18 of the 28 items require skill *x* and the rest do not, the 18 items will be coded as 1 and the remaining 10 items will be coded as 0 (see Figure 3). In practice, the TBR analyses normally involve more predictor variables, which make the evaluation of each split quite intensive. To handle this problem, Breiman et al.'s (1984) recursive partitioning algorithm is used to evaluate all possible splits of all possible predictor variables at each stage of analysis.

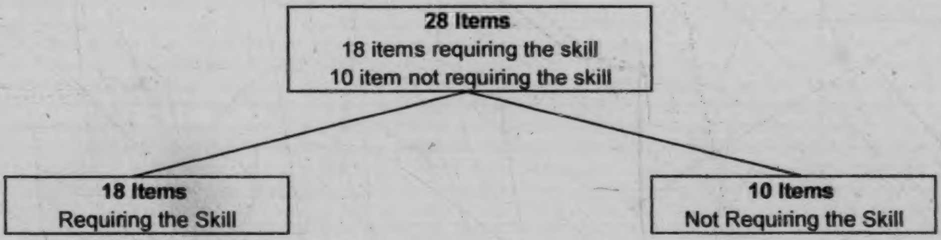

28 Items
18 items requiring the skill
10 item not requiring the skill
18 Items
Requiring the Skill
10 Items
Not Requiring the Skill

Figure 3. Illustration of the TBR splitting rules.

All possible splits are evaluated by deviance (i.e., the sum of squared differences between an observation and the expected value of all observations belonging to a single node). The best split is the one that maximizes the decrease in deviance between the parent node and the sum of the two child nodes. The deviance of the parent node, D (y, $\hat{y}$), is the sum of the deviances of all its members:

$$D(y,\hat{y}) = \sum (y_i - \hat{y})^2 \ , \qquad (3)$$

where $\hat{y}$ is the mean value of the criterion based on all observations in the node. The deviance of a potential split (D) is the sum of the deviances in the two child nodes:

$$D_{split}(y,\hat{y}_L,\hat{y}_R) = D(y,\hat{y}_L) + D(y,\hat{y}_R) = \sum_L (y_i - \hat{y}_L)^2 + \sum_R (y_i - \hat{y}_R)^2 \ , \qquad (4)$$

where $\hat{y}_L$ is the mean value of the criterion in the left child node and $\hat{y}_R$ is the mean value of the criterion in the right child node. The best split (ΔD) maximizes the deviance between the parent node and the two potential child nodes, and is calculated as,

$$\Delta D = D(y, \hat{y}) - D_{split}(y, \hat{y}_{L,} \hat{y}_R) \ . \tag{5}$$

Graphically, the results of the TBR analysis of test items can be depicted as a tree-like model of item clusters, where the horizontal location represents the predicted values for the criterion and the vertical location represents the percentage of variance in the criterion explained by the predictors from the best split (see Figure 4). As seen in the figure, at the

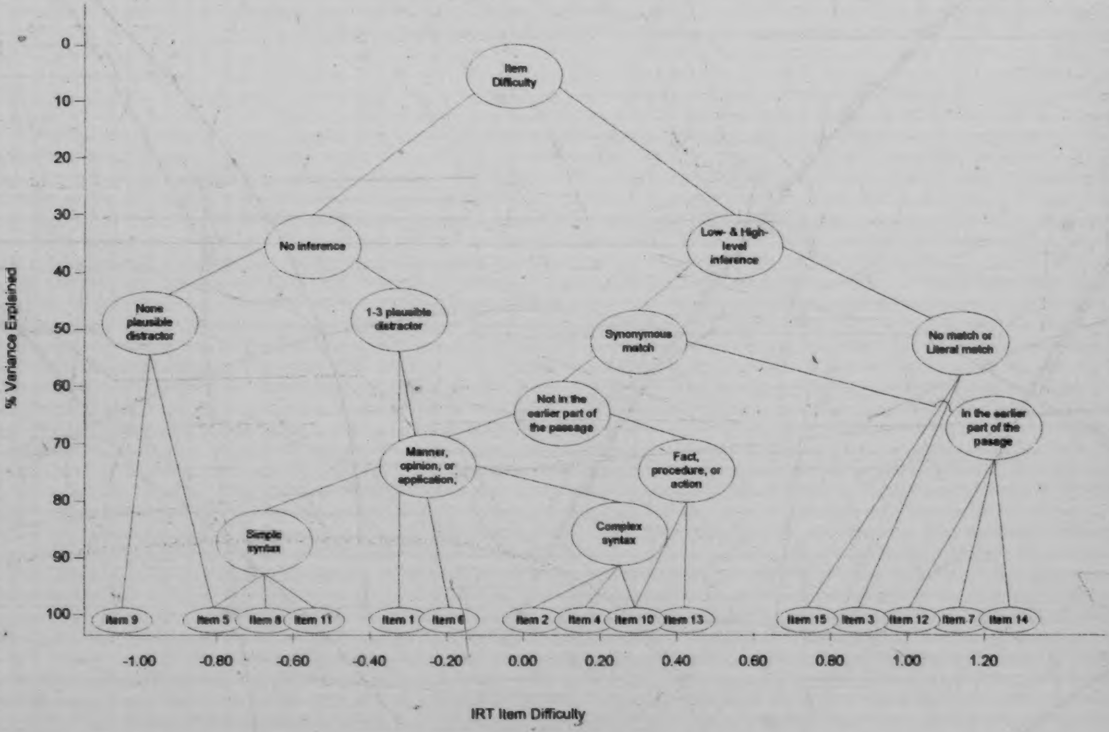

% Variance Explained
0
10
20
30
40
50
60
70
80
90
100
Item Difficulty
No inference
Low- & High-level inference
None plausible distractor
1-3 plausible distractor
Synonymous match
No match or Literal match
Not in the earlier part of the passage
In the earlier part of the passage
Manner, opinion, or application,
Fact, procedure, or action
Simple syntax
Complex syntax
Item 9
Item 5
Item 8
Item 11
Item 1
Item 6
Item 2
Item 4
Item 10
Item 13
Item 15
Item 3
Item 12
Item 7
Item 14
-1.00
-0.80
-0.60
-0.40
-0.20
0.00
0.20
0.40
0.60
0.80
1.00
1.20
IRT Item Difficulty

Figure 4. Illustration of the Tree-Based Regression Analysis.

top of the tree (parent node), all items are classified into a single cluster and 0% of the variance in the criterion is explained. That is, it is assumed that, at this node level, all items require a single, undifferentiated skill. At the bottom of the tree, each single item is classified into a unique cluster and 100% of the variance is explained. That is, it is assumed that, at this node level, each item tests a unique combination of cognitive structures. The best clustering solution (i.e., the optimal solution in terms of diagnosis and interpretative value) is the one by which the clusters defined by cognitive demands accounts for the maximum possible amount of the observed variation in the criterion. The final nodes that have no child nodes are called the terminal nodes.

After the TBR model is developed, pruning is often followed to increase the parsimony and interpretability of the model. Pruning is a process in which pairs of terminal nodes with common parents are collapsed by removing a split at the bottom of the tree. Pruning provides a useful method for model evaluation and model selection. For example, in item modeling where skills are difficult to code, pruning may be used to evaluate the effect of collapsing the terminal nodes associated with these skills.

The use of the TBR methodology for item modeling has several advantages. First, items selected from different test forms can be modeled simultaneously. For example, when the TBR is used for developing item difficulty models, the combinations of skills associated with differences in examinees' performances are determined by IRT item difficulty parameters. Hence, "no matter how many items are administered to individual examinees on individual test forms, sufficient within-skill-area item representation can always be achieved by analyzing additional test forms" (Sheehan, 1997, p. 351). Second, as TBR is powerful at modeling nonlinear relationships, it is a promising statistical tool for modeling complex ways in which cognitive structures of test tasks or items interact with different task or item characteristics to affect task or item difficulty. For example, Sheehan's (1997) TBR item difficulty model showed that the skills associated with each item set explained variance in item difficulty to various degrees. Based on this finding, Sheehan commented, "the SAT reading comprehension data would not be well fit by a linear model which required each skill to have the same effect on item difficulty, regardless of the item's schema classification" (p. 341). Third, as the TBR is powerful at classifying observations (items) into homogeneous groups rather than predicting values for the criterion (item difficulty), item modeling solutions from a TBR analysis are easier to interpret than those from a linear regression analysis (Breiman et al., 1984). Despite these advantages, detailed distinctions of item features (predictor variables) require more items in each category and extensive expert resources for item coding. Hence, while finer distinctions among the items would increase the variance explained and interpretability of the resulting models, it is often practically unfeasible (Huff, 2003).

Applications of the TBR to Item Modeling

Since Sheehan (1997) introduced the TBR methodology to item difficulty modeling, the TBR has been used for modeling item performance for the purpose of diagnostic score reporting (e.g., Huff, 2003; Sheehan, 1997; Wainer, et al., 2000), item development (e.g., Enright, et al., 2002), construct validation (e.g., Ewing & Huff, 2004; Rupp, et al., 2001; Sheehan & Ginther, 2001), and domain theory development (Strong-Krause, 2001). Sheehan (1997) used TBR to model item difficulty based on item processing characteristics in order to develop student- and group-level diagnostic feedback. He analyzed examinee responses to 78 verbal items on the SAT Verbal Reasoning Test. In his TBR analysis, the criterion was the 3-PL IRT item difficulty estimates, and the predictors were hypothesized skills required for item solution. Using a user-specified split, the items were first classified according to four processing strategies specified in Kirsch and Mosenthal (1990): Vocabulary, Main Idea and Explicit Statement, Inference, and Application or Extrapolation. The first split explained 20% of the observed variance in item difficulty. To explain more variance, each strategy node was split into two child nodes based on different skills within each strategy. For example, the Vocabulary strategy was further divided into Standard Word Usage and Poetic/Unusual Word Usage, and the Inference strategy was further divided into Specific Purpose and Attitude. This split explained about 50% more of the observed variance in item difficulty. Important variables in explaining item difficulty included vocabulary in context, complexity (i.e., gist or detail, concrete or abstract, explicit or implicit), cognitive operations required to arrive at the correct answer, and features of correct option and distractors.

In a subsequent study, Sheehan and Ginther (2001) successfully applied the TBR to the development of an item difficulty model for the Main Idea type reading items on the *TOEFL 2000*, based on cognitive processing features of the items. They coded the Main Idea items with three variables describing item-passage overlap features: Correspondence between correct response and textual information (0 = No Inference, 1 = Low Level inference, and 2 = High Level Inference), Location of Relevant Information (1 = Early, 2

= Middle, 3 = Late; and 4 = Entire Passage), and Elaboration of Information (scored as the percent of text that must be processed to correctly answer the item). The resulting cognitive processing model accounted for 87% of the variance in item difficulty, with Correspondence as the strongest predictor and Elaboration an insignificant predictor.

Rupp et al. (2001) modeled item difficulty of reading and listening comprehension items included in an ESL test using multiple regression and TBR analyses. Despite a small sample size (84 non-native English speakers of varying ability levels), two strengths are unique to their study. First, the combination of two techniques, multiple regression and TBR, provided multiple perspectives to more fully interpret the item difficulty models. Second, like previous applications of TBR to item modeling, the predictors in Rupp et al.'s models were cognitive demands of test items. However, Rupp et al. clearly defined three types of predictors associated with the cognitive processing underlying item performance: text characteristics (e.g., word count, sentence length, and information density), item characteristics (e.g., lexical overlap between correct answer and distractors), and text-by-item interactions (e.g., type of match). A limitation with their study might be the lack of strong evidence for combining the items across the modalities (reading and listening) in item modeling. Rupp et al. assumed that items could be grouped according to information processing characteristics common to both modalities. A think-aloud or dimensionality analysis may help clarify whether modeling item difficulty separately for reading and listening item groups would be better in terms of interpretability of the models.

Huff (2003) used the TBR to model item difficulty of the listening and reading items included in the *new TOEFL* to provide descriptive score reports regarding examinees' English language proficiency. In her study, the data were examinee performances on the Listening and Reading items from two parallel forms (1,372 examinees for Form 1 and 1,331 for Form 2). Her final models explained 56.0% of the variance in item difficulty for reading items and 48.0% for listening items. Several features distinguished her TBR analysis from previous TBR studies. First, both dichotomously- and polytomously-scored items were involved. Item difficulty parameters were estimated using the 3-PL IRT model for dichotomous items and the graded response model (Samejima, 1997) for polytomous items. Second, unlike previous TBR studies where items were classified using user specifications or expert coding of item features, Huff introduced cluster and dimensionality analyses to complement the subjective judgment of item classifications. Her study showed that dimensionality analyses facilitated item feature identification, item grouping, and substantive interpretations of item modeling solutions. Third, the predictors used in her TBR analyses were the existing item and passage codes developed by the TOEFL developers. These predictors included item and text characteristics, and were defined using Bachman's (1990) framework of communicative language ability and Mislevy's (1994) framework of evidence-centered design. However, as these existing codes were not defined specifically for the item difficulty research, sources of reading and listening item difficulty might not have been taken into account. Moreover, the interaction between item and text, that is, what an examinee is required to do and the type of information in the text, was not fully represented in the predictors used in her study.

The studies reviewed above reveal that defining item features is the fundamental issue in applying the TBR to item modeling, as what item features are included in the model and how they are coded are closely related to model interpretability (Ewing & Huff, 2004; Huff, 2003). In reading assessment, assessing examinees' processes when they read texts and respond to test items has been increasingly emphasized, and the methods in cognitive psychology such as task analysis and verbal reports have been used to gain insights into examinees' processes during task performance (Alderson, 2000). Accordingly, identifying cognitive processes underlying reading item performance needs to consider theoretical information, cognitive structures of items, and examinees' rendition of their actual item solving processes.

Literature Summary

This chapter reviewed the literature relevant to the cognitive-psychometric modeling of the MELAB reading items. The chapter first reviewed theories and models of the L2 reading and test-taking processes to inform the components in the cognitive model underlying the MELAB reading item performance. Then, the literature pertaining to the L2 reading assessment was reviewed to clarify the constructs assessed by the MELAB reading. Next, empirical studies of the factors affecting the L2 reading test item performance were critically reviewed to obtain scales that could be used to score the processing difficulty of reading test items and to inform the methods for the model development and validation. Finally, the TBR was reviewed in the context of modeling item performance as a way to empirically validate the cognitive model.

The literature discussed in this chapter has two major implications for the current study. First, a review of the theories of and research into the L2 reading and test-taking processes suggested a cognitive processing model hypothesized to underlie the MELAB reading test item performance. This model is consistent with current theories of L2 reading and its assessment (e.g., Alderson, 2000; Hudson, 1996) and the theoretical framework of CLA for language testing (Bachman, 1990; Bachman & Palmer, 1996), and takes into consideration the constructs assessed by the MELAB reading (ELI, 2003). Moreover, a review of the research into the factors affecting the L2 reading test item difficulty suggested scales that can be used to score the processing sources of cognitive complexity of the reading test items in the current study. These variables emerged from the cognitive processing model hypothesized to underlie the MELAB reading test item performance and linked examinees' cognitive processes with test item characteristics.

Second, a review of the studies modeling the L2 reading item performance framed a unified procedure to develop and validate the cognitive processing model underlying the MELAB reading item performance (see Figure 5). As can be seen in Figure 5, the model was developed and tested through four stages. First, models of the L2 reading and reading test taking processes, constructs of the L2 reading ability, and factors affecting the L2 reading performance were reviewed. Based on the theoretical information, an initial cognitive model was developed. Then, this model was empirically tested using a three-staged procedure: analyzing cognitive demands of the test items by raters, collecting students' verbal protocols of the processes they used to arrive at the correct responses, and examining the relationship between the proposed cognitive processes and item difficulty estimates using the TBR, a cognitively-based measurement model.

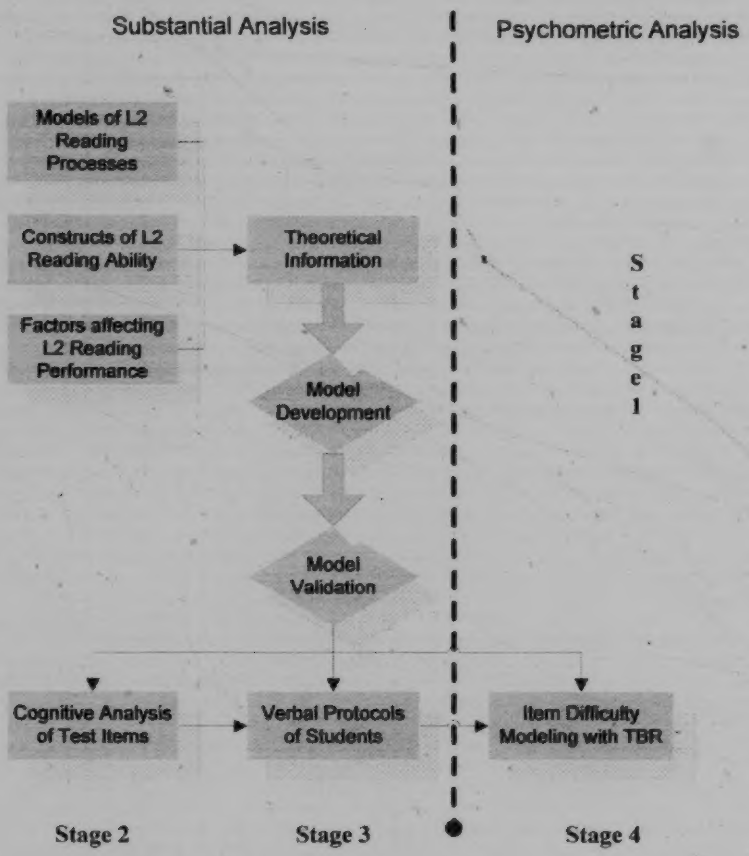

Substantial Analysis
Psychometric Analysis
Models of L2 Reading Processes
Constructs of L2 Reading Ability
Factors affecting L2 Reading Performance
Theoretical Information
Model Development
Model Validation
Stage 1
Cognitive Analysis of Test Items
Verbal Protocols of Students
Item Difficulty Modeling with TBR
Stage 2
Stage 3
Stage 4

Figure 5. Model Development and Testing Procedures.

Previous studies identifying the cognitive structures of the L2 reading test items either depended on theoretical information obtained from the literature, or on 'expert' analysis of the test items, or on examinees' reports of the processes that they used to respond to the test items. However, few studies have combined the three sources of data to determine the cognitive structures of test items. Further, no studies have modeled the

L2 reading test item performance using a combination of cognitive analyses of test items and advanced measurement models. In the current study, a review of the theoretical information regarding the L2 reading and its assessment provides the theoretical foundation for the proposed model. Systematic and reliable analysis of the cognitive demands of the MELAB reading test items refines the model and anticipates the cognitive processes used by examinees for item solving. Students' verbal reports of the cognitive processes used when they answer the MELAB reading items further refine the model and empirically validate the cognitive analysis of the test items by raters. Finally, relating the processing components in the cognitive model to empirical indicators of item difficulty using the advanced measurement model, TBR, further validates the cognitive model. To conclude, the present study is justified in employing a combination of theoretical information, cognitive analysis of test items by raters, students' verbal reports, and an advanced measurement model in an effort to develop a theoretically and empirically supported cognitive processing model underlying the MELAB reading test item performance. The initial cognitive processing model obtained from an analysis of the literature is outlined in the next chapter. The methods used to refine and validate this initial model and the findings obtained for each stage are described in Chapter 4, Chapter 5, and Chapter 6, respectively.

CHAPTER 3: AN INITIAL COGNITIVE PROCESSING MODEL

Following an analysis of the literature, a theoretically supported cognitive processing model was developed that was intended to explain the difficulty of the MELAB reading test items. The model contained three general cognitive categories. The first contained the reading processes assessed by the MELAB reading test items. The second included test management processes that examinees might use to arrive at their answers to the multiple-choice reading test items. The third category included testwise processes that examinees might use when deriving their answers. For each process, potential sources of processing difficulty were identified. Variables were then defined to score the cognitive complexity of these sources on the basis of theoretical and empirical relationships informed by previous research. The processing components in each of the categories and the variables related to these sources of processing difficulty are specified in this chapter.

Cognitive Processes and Cognitive Variables

Reading Processes Assessed by the MELAB Reading

Word recognition. Recognizing words and word meanings has been considered central to fluent L2 reading. Unfamiliarity with word pronunciation and scripts can lead to difficulty in processing letters in words and word meanings, which may in turn affect reading success of L2 readers (Alderson, 2000). Word recognition was included in the initial cognitive model to represent the process of (1) identifying words and word meanings using phonological and orthographic knowledge, and (2) understanding the meaning of a specific word or phrase in context. According to Urquhart and Weir (1998), reading test items that require examinees to identify words using advanced phonological and orthographical knowledge or to identify the meaning of an unknown word with few context clues are generally difficult to process for adult L2 readers. For example, the word *minute* can be pronounced as [mai'nju:t] or [minit]; *supercritical* can be recognized as the prefix *super* plus the root *critical*; *run* can have different meanings in different contexts. Consequently, word recognition was included as a variable in the initial cognitive model. It was hypothesized that processing difficulty would increase if advanced phonological and/or orthographic knowledge was required for identifying words or if few context clues were available for identifying the meaning of an unknown word. Following Bachman et al., (1995), the "Word Recognition" variable was coded as the degree to which examinees need to identify words using phonological or orthographic knowledge, or to understand the meaning of a specific word or phrase in context to correctly answer an item. The variable was measured using a 3-point scale: 0 = Word recognition is not required to successfully complete the item; 1 = Word recognition is somewhat involved, but not critical to the successful completion of the item; and 2 = Word recognition is critical to successful completion of the item. More specifically, the variable was coded 0 if examinees did not need to identify words using advanced phonological or orthographical knowledge or to identify the meaning of an unknown word in context; coded 1 if examinees need to identify words using more advanced phonological and/or orthographical knowledge, or need to identify the meaning of an unknown word in context, but such processes were not critical to correctly answering the item; coded 2 if examinees need to identify words using advanced phonological and/or orthographical knowledge, or to identify the meaning of an unknown word with few context cues, and such processes were critical to correctly answering the item.

Vocabulary knowledge. Vocabulary knowledge has been found to be critical for fluent reading by adult L2 readers (Alderson, 2000). When reading academic texts, a major source of processing difficulty is the lack of familiarity with infrequently used vocabulary (i.e., low frequency in everyday use), specialized vocabulary (i.e., jargon, academic, or technical words or phrases specific to the general topic of the text), or both (Carr, 2003; Hudson, 1996). Texts containing infrequently used and/or specialized vocabulary would increase the demands on decoding, understanding, and recall of the text (Bachman et al., 1995; Bachman et al., 1996; Carr, 2003), and this increased demand would increase item difficulty (Gorin, 2002). Vocabulary knowledge was included in the initial cognitive model to represent the process of answering an item through reading an academic text or part(s) of the text that contained a great deal of infrequently used and/or specialized vocabulary. It was hypothesized that texts containing more infrequently used and/or specialized vocabulary would be more difficult to process and decode for later use when responding to the items related to the texts. Again, based on Bachman et al.'s (1995) work, vocabulary knowledge was coded as the degree to which examinees need the knowledge of infrequently used and/or specialized vocabulary to correctly answer an item: 0 = knowledge of infrequently used or specialized vocabulary is not required to correctly answer the item; 1 = knowledge of infrequently used or specialized vocabulary is somewhat involved, but not critical to correctly answering the item; and 2 = knowledge of infrequently used or specialized vocabulary is critical to correctly answering the item. More specifically, the variable was coded 0 if the information requested by an item contained no infrequently used or specialized vocabulary; coded 1 if the information requested by an item contained some infrequently used or specialized vocabulary, but knowledge of such vocabulary was not critical to arrive at the correct response to the item; and coded 2 if the information requested by an item contained a great deal of infrequently used or specialized vocabulary, and knowledge of such vocabulary was critical to arrive at the correct response to the item.

Syntactic knowledge. Syntactic knowledge is required to construct coherent representations of sentence structures (Koda, 2005) and has been identified as a potential source of processing difficulty for L2 readers (Bachman et al., 1996; Alderson, 2000). Syntactic knowledge was included in the initial cognitive model to represent the process of understanding the relationship between ideas within a sentence using one's knowledge of syntax, grammar, punctuation, and/or parts of speech. It was hypothesized that reading test items requiring the processing of complex or infrequently used sentence structure, grammar, punctuation, or parts of speech would be more difficult than items requiring the processing of simple frequently used sentence structure, grammar, punctuation, and parts of speech. Syntactic knowledge was coded as the degree to which examinees required it to correctly answer an item, and was measured using Bachman et al.'s (1995) 3-point scale: 0 = Syntactic knowledge is not required to correctly answer the item; 1 = Syntactic knowledge is somewhat involved, but not critical to correctly answering the item; 2 = Syntactic knowledge is critical to correctly answering the item. More specifically, the variable was coded 0 if the information requested by an item contained simple, frequently used sentence structure, grammar, punctuation, and parts of speech; coded 1 if the information requested by an item contained more complex or less frequently used sentence structure, grammar, punctuation, or parts of speech, but knowledge of them was not critical to successful completion of the item; and coded 2 if the information requested by an item contained complex or infrequently used sentence structure, grammar, punctuation, or parts of speech, and knowledge of them was critical to successful completion of the item.

Knowledge of discourse structure. Knowledge of discourse structure is a critical element in L2 academic reading (Grabe, 2004, Hudson, 1996) and has been identified as a source of item difficulty on a variety of the L2 reading tests (Carr, 2003; Freedle & Kostin, 1993; Jamieson et al., 2000). Knowledge of discourse structure was included in the initial cognitive model to represent the process of understanding the relationship between sentences and text organization using cohesion, rhetorical organization, and information flow of the text. It was hypothesized that items requiring the reader to process texts with complex, infrequently used discourse structure would be more difficult than items requiring the reader to process texts with simple frequently used discourse structure. Knowledge of discourse structure was coded in terms of the degree to which examinees need knowledge of discourse structure to correctly answer an item and, following Bachman et al. (1995), was coded: 0 = Knowledge of discourse structure is not required to correctly answer the item; 1 = Knowledge of discourse structure is somewhat involved, but not critical to correctly answering the item; and 2 = Knowledge of discourse structure is critical to correctly answering the item. More specifically, the variable was coded 0 if the information requested by an item contained simple, frequently used discourse structure; coded 1 if the information requested by an item contained more complex or less frequently used discourse structure, but knowledge of them was not critical to successful completion of the item; and coded 2 if the information requested by an item contained complex or infrequently used discourse structure, and knowledge of them was critical to successful completion of the item.

Synthesis. Synthesis plays an important part in L2 academic reading (Hudson, 1996) and has long been a major skill measured on L2 reading tests (Alderson, 2000; Lunzer & Gardner, 1979). Synthesis was included in the initial cognitive model to represent the process of working across multiple places in the text to generate an organizing frame that was not explicitly stated in the text. It was hypothesized that items became more difficult as the level of synthesis increased, because integrating the information presented in different sentences or parts of the text requires more complex processing strategies than processing the information contained within a single word, phrase, or sentence. Based on Kirsch and Mosenthal's (1990) taxonomy of levels of synthesis, the "Synthesis" variable was coded as the degree to which examinees need synthesis to correctly answer an item, and was rated using a 3-point scale: 0 = No synthesis is required to correctly answer the item, 1 = Low-level synthesis is required to correctly answer the item, and 2 = High-level synthesis is required to correctly answer the item. More specifically, the variable was coded 0 if the information requested by an item was contained within a single place; coded 1 if the information requested by an item was contained within multiple adjacent sentences; and coded 2 if the information requested by an item was contained within multiple nonadjacent sentences or diffused across the passage.

Drawing inferences. Drawing inferences is part of the reading process of adult L2 readers (Alderson, 2000) and a major skill measured on the L2 reading tests (ELI, 2003; Grabe, 2004). Prior research revealed that the level of inference required to arrive at the correct response contributes significantly to reading test item difficulty (Davey, 1988; Embretson & Wetzel, 1987; Gorin, 2002; Kasai, 1997; Rupp et al., 2001; Sheehan & Ginther, 2001). Inference was included in the initial cognitive model to represent the process of drawing inferences and conclusions based on information presented in the text. According to Rupp et al. (2001), items become more difficult as examinees need higher levels of inference to solve an item, because making inferences is more cognitively demanding than recognizing explicitly stated information. The "Inference" variable was coded as the level of inference that examinees need to correctly answer an item, and was measured using a 3-point scale: 0 = No inference is required to correctly answer the item; 1 = Low-level inference is required to correctly answer the item; and 2 = High-level inference is required to correctly answer the item (Embretson & Wetzel, 1987; Gorin, 2002; Sheehan & Ginther, 2001). More specifically, the variable was coded 0 if the information requested to correctly answer an item was explicitly stated in the passage, if examinees could answer the item without reference to the passage, or if the relationship of item to passage was unclear; coded 1 if the information requested to correctly answer an item was implicitly presented in a specific part or several specific parts of the passage; and coded 2 if the information requested to correctly answer an item could be generated based only on the examinee's global understanding of the entire passage, capability of predicting the continuation of arguments or events, or ability to relate information in the passage to the real world.

Pragmatic knowledge. Analyzing pragmatic or rhetorical purposes of the text creator is a crucial component in L2 reading process (Bachman et al., 1995; Enright et al., 2000). Further, the ability to understand authors' pragmatic and rhetorical purposes has been a major construct assessed by L2 reading tests such as TOEFL reading and MELAB reading (ELI, 2003; Enright et al., 2000; Jamieson et al., 2000; Vanderveen, 2004). This important reading skill, termed as Pragmatic Knowledge, was included in the initial cognitive model to represent the process of analyzing authors' pragmatic or rhetorical purposes. In reading assessments, pragmatic and rhetorical purposes of the text creator has been proposed as a factor affecting reading item difficulty (Enright et al., 2000; Jamieson et al., 2000; Kirsch & Mosenthal, 1990; Rupp et al., 2001). For example, it has been argued that factual information with the primary purpose to inform the reader would be easier to process, understand, and recall than counterfactual information with the primary purpose to persuade the reader, and this effect should be reflected by item difficulty (Bachman et al., 1995; Enright et al., 2000; Jamieson et al., 2000; Rupp et al., 2001). Based on Jamieson et al.'s (2000) description of task factors that can account for reading test item difficulty, pragmatic knowledge was coded as authors' pragmatic or rhetoric purposes for specific part(s) of the text that examinees need to understand to correctly answer an item. Pragmatic knowledge was coded using a 5-point scale with higher numbers representing the purposes that require more complex processing strategies to distinguish or understand and thus leading to increased item difficulty: 1 = The information requested by an item was intended to inform a fact; 2 = The information requested by an item was intended to state a procedure or to describe an action; 3 = The information requested by an item was intended to analyze manner or goals; 4 = The information requested by an item was intended to express authors' or others' attitudes or opinions, to explain cause or effect, to provide evidence, to support a position, or to persuade the reader; and 5 = The information requested by an item was intended to establish equivalence or difference, to generate a theme from the information provided in the passage, or to apply the information provided in the passage to the real world.

Test Management Processes

Prior research has identified a series of processes that examinees use to manage their solutions to the multiple-choice reading test items. In the present study, these processes were called test management processes. The four test management processes included in the initial cognitive model are presented next.

Locating specific information requested by item. Locating specific details in the text is a test management process used by adult L2 readers when responding to reading test items (Abbott, 2005; Alderson & Lukmani, 1989; Enright et al. 2000; Jamieson et al., 2000). The location of the requested information has been found to be a potential source of reading item difficulty when identifying the information requested by an item in the text passage. For example, requested information located earlier in the text may no longer be in the reader's short-term memory, thereby increasing the number and complexity of cognitive operations the examinee needs to use to answer the item (Sheehen & Ginther, 2001; Rupp et al., 2001). Locating specific details in the text was included in the initial cognitive model to represent a test management process. The variable "Location of Information" was coded by dividing the entire passage into three equal sections based on word count, with "1" being the later part (i.e., the third section) of the passage which was taken to be the most recent in memory, "2" being the middle part of the passage (i.e, the second section) which was taken to be further in memory, and "3" being the earlier part of the passage (i.e., the first section) which was taken to be out of readers'short-term memory. Then, the section in which the requested information could be found was recorded (Rupp et al., 2001). Based on Rupp et al. (2001) and Sheehan and Ginther (2001), this variable was rated using a 4-point scale: 1 = The information requested to correctly answer an item is located in the third section of the passage, 2 = The information requested to correctly answer an item is located in the second section of the passage, 3 = The information requested to correctly answer an item is located in the first section of the passage, and 4 = The information requested to correctly answer an item is located across the entire passage or beyond the passage. It was assumed that the cognitive processes required for each code on this scale were progressively more complex, which would lead to increased item difficulty.

Matching information in question to information in text. During taking a reading test, examinees may match the information given in a question to the information provided in the text (Abbott, 2005; Cohen & Upton, 2006). Prior research revealed that the process of matching question to text consisted of a range of strategies that varied in difficulty and a variety of conditions that rendered processing strategies more or less difficult (Anderson, 1982; Embretson & Wetzel, 1987; Freedle & Kostin, 1993; Gorin, 2002; Huff, 2003; Kirsch & Mosenthal, 1990; Jamieson et al., 2000; Rupp et al., 2001; Sheehan & Ginther, 2001). Matching was included in the initial cognitive model to represent the processes that examinees used to match the information given in a question stem to the corresponding information in a text. Depending on the correspondence between the phrasing used in the question stem and the phrasing used in the text, the matching process was considered to involve examinees' directly matching the key vocabulary in the question stem to the key vocabulary in the text, and identifying or formulating a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence in the text. It was hypothesized that question stems that required examinees to perform a literal or verbatim match to the text would be easier than question stems that required examinees to perform a paraphrase or synonymous match to the text, which would be easier than question stems that could not be matched to the text. Question stems having a literal match to the text were assumed to be easy because the requested information could be directly identified in the text without any transformation being conducted. Question stems having a synonymous match to the text were assumed to be more difficult, because wording and order of the information in both question and text must be transformed to map the question to the information requested to correctly answer the item. Question stems that could not be matched to the text were assumed to be the most difficult, because the idea structure of the entire passage had to be reworded, reordered, and integrated in order to infer the information requested to correctly answer the item (Gorin, 2002). The variable "Type of Match" was coded for the degree of correspondence between the phrasing used in the question stem and the phrasing used in the text. Type of Match was rated using a 3-point scale with higher numbers representing more complex encoding comparisons that examinees need to answer an item correctly: 1 = One or more words or phases used in the question stem exactly match the word(s) or phrase(s) used in the text (i.e., literal match); 2 = the wording used in the question stem is the synonym or paraphrase of the wording used in the text (i.e., synonymous match); 3 = The wording used in the question stem can not be matched to the wording used in the text (i.e., no match).

Evaluating alternative choices. Before selecting the answer, examinees may evaluate the alternative choices to see which one best fits the syntactic and semantic features of the question and the passage information (Drum et al., 1981). Plausibility of the alternative choices has been found to be a major predictor for reading item difficulty (Drum et al., 1981; Jamieson et al., 2000; Rupp et al., 2001). Evaluating alternative choices was included in the initial cognitive model to represent the processes of selecting the one option that best fits the requirements of the question and the idea structure of the text and eliminating the choice(s) that appear unreasonable based on paragraph or overall passage meaning. It was hypothesized that item difficulty would increase as the number of plausible distractors increased, as examinees would need to make finer discriminations among the options to identify the correct answer. The number of plausible distractors was counted based on how many of the distractors had lexical overlap with or the same idea structure (both explicitly or implicitly) as the text (Rupp et al., 2001). Since each item on the MELAB had four options and a single key, the variable "Number of Plausible Distractors" was coded: 0 = No plausible distractors; 1 = One plausible distractor; 2 = Two plausible distractors; 3 = Three plausible distractors.

Using topical knowledge. Examinees may answer an item through using their knowledge related to the content or subject matter of the text (Carr, 2003). While the MELAB reading test is not intended to measure examinees' topical knowledge, having knowledge related to the topic or content of the text facilitates examinee performance (Alderson, 2000). Using topical knowledge was included in the initial cognitive model to represent the process of constructing situation models using one's knowledge related to the content or subject matter of the text. It was hypothesized that items based on text with unfamiliar topics would be more difficult for examinees than items based on text with familiar, everyday topics, since examinees may need special knowledge related to text content. Topical knowledge was coded as the degree to which examinees need topical knowledge to correctly answer an item and, following Bachman et al. (1995), was rated using a 3-point scale: 0 = Topical knowledge is not required to correctly answer the item; 1 = Topical knowledge is somewhat involved, but not critical to correctly answering the item; 2 = Topical knowledge is critical to correctly answering the item. More specifically, the variable was coded 0 if the item was based on a text with a familiar everyday topic; coded as 1 if the item was based on a text with a less familiar topic related to a specific subject area, but familiarity with the topic was not critical to correctly answering the item; coded as 2 if the item was based on a text with an unfamiliar topic related to a specific subject area and familiarity with the topic was critical to correctly answering the item.

Testwise Processes

In addition to the reading and test management processes, examinees may correctly answer multiple-choice test items using testwiseness (Cohen & Upton, 2006; Gao, 2002; Powers & Leung, 1995). Following Cohen and Upton's (2006) taxonomy of testwise processes that examinees may use to arrive at answers to the new TOEFL reading test items, three testwiseness elements were included in the initial cognitive model. The first was to use cues in the other items sharing a common passage to answer an item under consideration. For example, examinees may use the question stem of one item to identify the key to another item. Examinees may also use the interrelatedness of the items sharing the same passage to construct a situation model that can be used to reason what the correct response is. The second was to select an option through common sense, vague hunches, or intuition. The third was to eliminate or choose an option based on surface features of answer choices. For example, examinees may select an answer based on the observation that one option is longer than other options. In the current study, testwiseness was considered irrelevant to the constructs assessed by the MELAB reading and thus a potential threat to test validity. It was assumed that an item would be easier if examinees could employ testwiseness elements to correctly answer an item. The three testwise processes were coded as whether examinees could use the testwiseness elements considered to arrive at the correct answer: 0 = No, 1 = Yes.

A Summary of the Initial Model

Table 1 at the end of this chapter summarizes the components of the initial cognitive processing model, sources of processing difficulty for L2 reading test items, and variables used to score the cognitive complexity of each process. Column 1 presents the 14 components of the cognitive processing model. Column 2 describes the major cognitive processes associated with each component. Column 3 explains sources of processing difficulty, which links the processing sources of cognitive complexity to reading item difficulty. Column 4 presents example situations where different levels of processing difficulty are involved. The last column presents the scales used to score the processing difficulty of reading test items. For each scale, higher scores correspond to higher processing difficulty thus leading to more difficult items.

Take the first component, word recognition, as an example. The major cognitive processes associated with this component include recognizing words using phonological and orthographic knowledge, and identifying meaning of unknown words in context. The major sources of processing difficulty in the process of word recognition are the sound and orthography of words and the context of unknown words (Bachman & Palmer, 1996; Urquhart & Weir, 1998). Words with different sounds and complex orthography increase the demands on decoding and understanding, which should be reflected in increased reading item difficulty. For example, the word "*minute*" must be pronounced as [mai'nju:t] to recognize that it means *tiny*. The word "*wind*" means *air movement* when pronounced as [wind], and means *to wrap something around a centre or to coil thread* when pronounced as [waind]. The word "counterintuitive" can be recognized as the prefix *counter* plus the root word *intuitive*. In addition to phonological and orthographical knowledge, context plays an important part in identifying meanings of a specific word. For example, the word "run" means *to compete in a race for elected office* in the context *ran for mayor*, means *to move or go quickly* in the context *ran for the police*, and means *to flow in a steady stream* in the context *water runs from the spring*. The word "magazine" means *a periodical containing a collection of articles, stories, and pictures*, or means *a compartment for bullets*. The word "house" means *a dwelling* as a noun and means *to store or shelter* as a verb. The variable used to score this processing source of cognitive complexity is the degree to which the processes of identifying words using advanced phonological and orthographical knowledge or identifying meanings of an unknown word in context is required to correctly answer an item, and was coded using a 3-point scale with 0 representing not really required, 1 representing somewhat involved, and 2 representing critically required to correctly answer the item.

Table 1

Cognitive Processes Used by L2 Readers When Answering the Multiple-Choice Reading Test Items

Code/ Process	**Definition**	*Processing difficulty increases if:*	**Example**	**Variable Name, Abbreviation, Description, & Scoring**
Reading Processes Assessed by the MELAB Reading Test Items				
R1 Word recognition	Identify words and word meanings using phonological and orthographic knowledge; Recognize and understand the meaning of a specific word or phrase using context clues (i.e., neighboring words, sentences, or overall passage)	Advanced phonological and/or orthographic knowledge is required; Less contextual info is available to help identify the meaning of an unknown word	*Minute* can be pronounced as [mai'nju:t] or [minit]; *supercritical* can be recognized as super + critical; *Run* can mean to flow in a steady stream (e.g., water runs from the spring), or to compete in a race for an elected office (e.g., run for mayor), depending on the context	Word Recognition Required (R1): Degree to which examinees need to identify words using phonological or orthographic knowledge, or to understand the meaning of a specific word or phrase in context to correctly answer the item 0 = Not required; 1 = Somewhat involved; 2 = Critical
R2 Using vocabulary knowledge	Understand academic texts with infrequently used vocabulary (i.e., low frequency in everyday use) and specialized vocabulary (i.e., jargons, academic, or technical words or phrases specific to the general topic of the text)	Texts contain infrequently used and/or specialized vocabulary	Infrequently used vocabulary: ebullient, cacophony, salad days; Specialized vocabulary: *deposition* in chemical engineering, *morpheme* in linguistics, *homoscedacity* in statistics	Vocabulary Knowledge Required (R2): Degree to which examinees need to understand texts with infrequently used and/or specialized vocabulary to correctly answer the item 0 = Not required; 1 = Somewhat involved; 2 = Critical

R3 Using syntactic knowledge	Understand the relationship of ideas within the sentence using knowledge of syntax, grammar, punctuation, or parts of speech	Understanding infrequent or complex sentence structure, grammar, punctuation, or parts of speech is required	Inversion of subject and verb; passive voice; two or more clauses connected by a subordinate conjunction, a relative pronoun, or a relative adverb	Syntactic Knowledge Required (R3): Degree to which examinees need to understand the relationship of ideas within the sentence using knowledge of syntax, grammar, punctuation, or parts of speech to correctly answer the item 0 = Not required; 1 = Somewhat involved; 2 = Critical
R4 Using knowledge of discourse structure	Understand the relationship between sentences and organization of the text or portion of the text using cohesion (i.e., cohesive devices used in text to indicate the relationship between text utterances or to replace previously occurring parts of the text), rhetorical organization (i.e., text rhetorical features and organization patterns), and information flow (i.e., idea structure of the text)	Understanding complex discourse structure is required	Cohesion: reference, substitution, ellipsis, and conjunction; Rhetorical organization: cause/effect, comparison/contrast general/specific, problem/solution, using illustrations, topic sentence and supporting details; Information flow: using introduction, topic sentence, illustrations, transitions, conclusion, definitions, classifications, and supporting details	Knowledge of Discourse Structure Required (R4): Degree to which examinees need to understand the relationship between sentences and organization of the text or portion of the text using cohesion, rhetorical organization, and information flow to correctly answer the item 0 = Not required; 1 = Somewhat involved; 2 = Critical
R5 Synthesis	Integrate, relate, or summarize the information presented in different sentences or parts of the text to generate an organizing frame that is not explicitly stated in the text	Working across multiple places in text is required, since the number and complexity of cognitive operations increase	No synthesis: the requested information is contained within a single word, phrase, or sentence Low-level synthesis: the requested information is contained within multiple adjacent sentences High-level synthesis: the requested info. is contained within multiple nonadjacent sentences or diffuses across the text	Synthesis Required (R5): Degree to which examinees need to work across multiple places in text to correctly answer the item 0 = No synthesis; 1 = Low level synthesis; 2 = High level synthesis

R6 Drawing inferences	Draw inferences and conclusions or form hypotheses based on information implicitly stated in the text	The requested info is implicitly given in the text, as relating the requested info to the info implicitly given in the text is more cognitively demanding than recognizing explicit info in the text	No inference: the requested info is explicitly stated in the passage, the item can be answered without reference to the passage, or the relationship of item to passage is unclear Low-level inference: the requested information is implicitly stated in certain part(s) of the passage High-level inference: the requested info can only be generated based on understanding of the entire passage, predicting the continuation of arguments or events, or relating info in passage to the real world	Inference Required (R6): Degree to which examinees need to draw inferences and conclusions based on information implicitly stated in the text to correctly answer the item 0 = No inference; 1 = Low level inference; 2 = High level inference
R7 Using pragmatic knowledge	Understand pragmatic and rhetorical purposes of the text creator	More complex processing strategies are required to distinguish or understand the purpose or intent of the text creator	Fact: When you are fully awake and alert, your EEGs contain many beta waves, relatively high-frequency, low-voltage activity (selected from Enright et al., 2000, p. 20) Q: What your EEGs contain when you are fully awake? Opinion: Perhaps mimeographed Christmas letters should be used as a vanity indicator, since they expose those among us who yielded to, rather than resisted, the pervasive temptation to blow one's own horn (selected from Enright et al., 2000, p. 21) Q: What is the author's position with regard to Christmas letters?	Purpose of Information (R7): Pragmatic or rhetorical purpose of the information requested to correctly answer the item 1 = to inform a fact; 2 = to state a procedure or to describe an action; 3 = to analyze manner or goals; 4 = to express author's or others' opinions, to explain cause or effect, to provide evidence, to support a position, or to persuade the reader; 5 = to find out equivalence or difference, to generate a theme, or to apply to the real world

Test Management Processes				
T1 Locating specific details in text	Locate specific information requested by question	The requested info. is located in the earlier part of the text, as it is no longer in one's short term memory	Items requiring information located in the earlier part of the passage would be more difficult to recall than items requiring information located in the later part of the passage; Items requiring information based on the entire passage or beyond passage do not require locating specific details but more complex processing strategies	Location of Information (T1): The section of the text (1st, 2nd, last section, or entire passage) in which the requested information can be found 1 = 3rd section of the passage; 2 = 2nd section of the passage; 3 = 1st section of the passage; 4 = Entire passage or beyond the passage
T2 Matching question to text	Match the information given in the question stem to the relevant info in the text: Match key vocabulary items given in a question stem to key vocabulary items in the text; identify or formulate a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence in the text	Matching is required and the lexical overlap between the text and the question is lower, as more complex processing strategies are required to identify or generate the requested information from the text	Questions containing a key word(s) that is the verbatim of the info in the text would be easier than questions that are paraphrases of the info in the text, *e.g.*, The question *What is the recommended adult dosage?* requires a literal match between the phrase given in the question stem, *adult dosage*, and the corresponding phrase in the text; Questions that do not need matching the textual information would require more complex processing strategies and thus leading to increased item difficulty	Type of Match (T2): Correspondence between the wording used in the question stem and the wording presented in the text 1 = Literal match; 2 = Synonymous match; 3 = No match
T3 Evaluating alternative choices	Select the one that best fits the requirements of the question and the idea structure of the text and eliminating the option(s) that appear unreasonable based on paragraph or overall passage meaning	The number of plausible distractors increases and finer discriminations of the options are required to identify the correct option	One or more distractors appear reasonable and difficult to deny by information in the text, e.g., distractors have lexical overlap with the text or correct option, distractors are plausible given the situation/idea structure described in the text	Number of Plausible Distractors (T3): Number of distractors having lexical overlap with the text or plausible given the situation described in the text 0 = None; 1 = One; 2 = Two; 3 = Three

T4 Using topical knowledge	Construct situation models using the knowledge directly relevant to the text topic	The topic of the text is arcane or in less familiar settings, as less familiar topics are harder to process and recall and prevent examinees from performing to the best of their ability	Easier topics: concrete, imaginable, interesting, or everyday topics; texts in familiar settings; texts describing real objects, events or activities Harder topics: texts on abstract, arcane topics or related to specific subject areas	Topical Knowledge Required (T4): Degree to which examinees need topical knowledge to correctly answer the item 0 = Not required; 1 = Somewhat involved; 2 = Critical
Testwiseness				
TW1 Using clues in other items	Use cues in the other items of the item set to answer the item under consideration	The interrelatedness of the items is lower; fewer cues in the other items sharing a common text	The answer to one item might be cued from the question stem or options of the other item(s) of the item set	Item Cues (TW1): Examinee can use cues in other items to correctly answer the item 0 = No; 1 = Yes
TW2 Guessing	Select an option through common sense or prior knowledge unrelated to the passage topic; use vague hunches or intuition; guess randomly among the choices	The chance of guessing to the correct answer is lower	Arrive at the correct answer using culture knowledge, tend to guess a particular choice (e.g., B or C), or the correct answer is the easiest one to understand	Guessing (TW2): Examinees can use common sense or prior knowledge unrelated to the passage topic, vague hunches or intuition correctly, or even random guessing to correctly answer the item 0 = No; 1 = Yes
TW3 Using surface features of answer choices	Eliminate or choose an option based on surface features of answer choices rather than on the textual information	Less cues from the answer choices	The correct answer is selected based on its length, tone, concreteness, clarity, wording, complexity, location, similarity to other choices or question stem	Surface Feature of Options (TW3): Examinees can use surface features of answer choices, such as length or tone to correctly answer the item 0 = No; 1 = Yes

CHAPTER 4: ITEM CODING BY RATERS

As mentioned earlier, after an initial cognitive processing model was developed based on the review of the pertinent literature, which has been addressed in Chapters 2 and 3, the initial model was validated using a three-stage procedure. In these three model-validation stages, two forms of the MELAB reading test were analyzed. The first stage of the model validation involved judges familiar with L2 reading and Mandarin-speaking students who take the MELAB tests. The judges identified the cognitive processes required to correctly answer each reading item on both forms. This analysis served to link the initial cognitive model presented in Chapter 3 to the MELAB reading items, to refine the initial cognitive model, and to verify the scoring scales presented in Table 1. In the second stage of the model validation, verbal report data were collected from advanced Mandarin-speaking L2 students as they worked through both forms of the MELAB reading test. The results were used to validate the rating results and to further refine the initial cognitive model so that the components of the model were clearly defined, informative, and faithful descriptors of examinees' cognitive processes underlying the MELAB reading test item performance. In the last stage of the model validation, the two revised cognitive models obtained at the end of the previous stage were empirically tested through item difficulty modeling using the tree-based regression, a cognitively-based measurement model. The combination of the three stages of analyses provided a unified sequential procedure to validate the proposed model, with each data analysis task reflecting part of the triangulation of data sources.

This chapter outlines the method used and presents the results for the first stage of the model validation — cognitive analysis of each item on both forms of the MELAB reading test from the perspective of raters with knowledge of the reading process and the population of Mandarin-speaking ESL university-level students. The methods and results for the second and third stage of the model validation are presented in the following two chapters, respectively.

Method

Raters

As this study focuses on the Chinese language group, three raters familiar with the reading process and the population of Mandarin-speaking students who use English in a college or university setting analyzed the cognitive processes required to correctly answer the MELAB reading test items. All raters were graduate students in educational psychology at the University of Alberta. Two of the raters were trained psychometrians with expertise in measurement and cognition and experience in test design and construction. The third rater was a trained applied linguist with expertise in L2 reading and experience in teaching reading to adult Chinese EFL learners. To recruit raters, ten days before the data collection, the researcher sought permission from the Department of Educational Psychology to send a recruiting letter (see Appendix A), via the educational psychology graduate student mailing list. Four students who responded and expressed an interest in participation were contacted by the researcher. One withdrew from the study before the formal rating procedure began; the remaining three participated in the rating.

Materials

MELAB test forms. All MELAB forms are designed to be parallel and are linked to a common scale. Each form consists of three parts: Part 1 (Composition), Part 2 (Listening), and Part 3 (Grammar, Cloze, Vocabulary, and Reading). For the purposes of this study, the analyses were performed using the reading passages and accompanying items included in Forms E and F administered during the year 2003-2004[11]. Each form contained 20 four-option multiple-choice test items based on the four passages, with five items per passage. The passages included in each form ranged from 229 to 265 words in length, had comparable readability levels as measured by standard readability formulas, and were on topics of social sciences, biological sciences, physical sciences, and agriculture (ELI, 2003).

Rating instrument. In order to enhance the rating reliability, a rating instrument was employed for the analysis of the cognitive processes required to correctly answer the MELAB reading test items. The rating instrument consisted of four parts. Part 1 was the initial cognitive model and cognitive variables described in Table 1 (see Chapter 3). The components of the initial cognitive model listed in the first column of Table 1 provided the framework for developing the rating instrument. The next three columns, that is, the definition of cognitive processes covered, explanation of rationale, and demonstration of example item features, clarified concepts involved in the rating process. The cognitive variables and variable scorings presented in the last column of Table 1 were provided to the raters to judge the cognitive processes called for by the MELAB reading test items.

Part 2 of the rating instrument contained the rating scales provided to the raters to code the reading items included in the two test forms, with the rows of the rating scale containing the items and the columns containing the 14 elements of the initial cognitive model. Part 3 of the rating instrument requested the raters to provide evidence for their ratings for the variables “Word Recognition Required” and “Vocabulary Knowledge Required”. For example, for the variable “Vocabulary Knowledge Required”, raters were asked to first code an item using the cognitive variables presented in the last column of Table 1 and the rating scale, and then to bracket the text that contained the information requested to correctly answer the item and list the words that they considered as infrequently used or specialized in that part of the text. Part 4 of the rating instrument allowed the raters to indicate any cognitive processes that were not included in the initial cognitive model but were required for solving the reading items included in Forms E and F. A copy of the full rating instrument with instructions is provided in Appendix B.

Data Collection

Pilot study. To test the rater training and rating procedures prior to their use, a pilot study for Phase 1 was conducted with two graduate students (one Chinese, one native-English speaker) in Educational Psychology at the University of Alberta. They were both familiar with the analysis of task characteristics for foreign language tests, and possessed expertise in L2 reading and experience in teaching reading English as a second or foreign language to L2 adults. The pilot study started with the training of the two raters. The initial cognitive model and the rating instrument were explained to the raters and discussed using the first passage and its associated items on Form E. As part of the training, the two raters coded the items related to this first passage and discussed each rating as it was completed with the researcher. Then, a copy of the researcher's rating of the items related to the second passage was provided to each pilot test rater for them to consult while they were rating the items connected to Passage 2. The items related to the remaining two passages on Form E were rated independently by the two raters. After they had completed these two passages, their rating outcomes were discussed with the researcher to reach consensus. Based on the pilot study, the rater training and rating procedures were refined and then used in the main study.

Main study. After signing a consent form (see Appendix C), the three raters separately rated the reading test items included in both forms using the rating instrument. In order to enhance the rating reliability, a one-hour group training session was held on day 1. During the training session, the researcher introduced the study and the MELAB reading test forms, acquainted the raters with the content of Table 1, described the rating form, and clarified the rating procedures. Discussions were encouraged as a way of achieving common definitions and understanding of the procedure. As part of the training session, all raters separately rated the first two passages with their associated items included in Form E for practice. The raters were asked to first answer the set of items. After completing the items, the answer key was provided for them to mark their answers. Then, the raters used the rating instrument to rate the items in terms of all possible cognitive processes required to correctly answer each item. Upon completion, the rating results for these two passages were discussed, the rationale for rating particular cognitive processes shared, and inconsistencies resolved.

After the training session, the raters independently rated the remaining 10 items included in Form E and the 20 items in Form F of the MELAB reading test on days 2, 3 and 4. To ensure that the procedure was followed exactly, each rater was provided with three envelopes, which contained instructions and materials for each step of the rating task. Envelope A contained Form E and Form F of the MELAB reading test. The raters were instructed to read the passages and answer the items as if they were indeed taking the reading test. Upon completion of this task, they were instructed to open Envelope B, which contained the answer keys (provided by the ELI) to the items included in both forms of the MELAB reading tests. The raters were instructed to check and correct their answers. Upon completing the marking task, they were instructed to open Envelope C, which contained the rating instrument and instructions for rating the items in terms of the cognitive processes required to correctly answer each item. The raters were asked to complete the entire task in three days, and to return their completed work with all the instructions and materials in the original envelopes to the researcher by 5pm of the third day. A copy of the final rating procedures with instructions for rater training and item rating is provided in Appendix D.

Data Analysis Procedures

Checking rating reliability. The researcher entered the rating data collected from the three raters into the Microsoft Excel 2000 (Microsoft Corporation, 2000) and verified the data entry to ensure 100% accuracy. Rater consistency was examined using G-theory (Brennan, 2004). G-theory was used for its ability to offer a more comprehensive framework for studying rater data and to examine rater performance across a number of different factors, such as cognitive processes and items. A fully crossed item-by-process-by-rater mixed effect G-study design was used. Items were treated as the object of measurement, raters were a random facet, and cognitive processing components a fixed facet. The computer program GENOVA (Crick & Brennan, 1983) was used to obtain the variance components and generalizability coefficient.

Reaching consensus ratings. The researcher summarized by hand the rating data provided by the three raters on the evening of day 4. Then, a working session involving the researcher and the three raters was held on day 5 for the raters to look at the rating summary and to reach consensus on the item ratings for which there was a lack of agreement. Following the meeting, the researcher entered the consensus ratings into the Microsoft Excel 2000 and verified the entered data for 100% accuracy. The final set of item ratings was formatted into two 20 x *k* matrices (20 is the total number of items on each test form and *k* is the cognitively-based item features), with one matrix for Form E and the other for Form F.

Results

Rating Reliability

The summary of the initial item ratings provided by the three raters is displayed in Appendix E. G-theory was applied to the data presented in Appendix E to check the rating reliability for all items rated by the three raters. Table 2 presents the variance components and generalizability coefficient obtained from the G-study and D-study.

Table 2

Variance Components and Generalizability Coefficient from the G-Study and D-Study

Source of Variability	Degree of Freedom	Variance Component	Percentage of Total Variance
Item	39	0.0295	3.85
Rater	2	0.0023	0.16
Process	13	0.4558	59.71
Item-Rater	78	0.0173	1.19
Item-Process	507	0.3071	20.78
Rater-Process	26	0.0073	0.50
Residual	1014	0.2413	13.82
Generalizability Coefficient ρ			0.90

Several notable findings can be observed from the table. First, the generalizability coefficient ρ was 0.90, which indicated that the items were fairly consistently rated by the three raters. Further examination of the relative sizes of the variance components involving raters provided a more comprehensive understanding of the raters' performance. The effects involving raters and the interaction of raters with item and process accounted for a negligible amount of the total variance. Only 0.16% of the total variance was accounted for by the rater effect, 1.19% by the item-rater interaction, and 0.50% by the rater-process interaction. Hence, it was concluded that the raters performed consistently across processes and across items. Second, the first and second largest variance components came from the process effect and the item-process interaction respectively, indicating that different processes were required to solve different items. Third, the item effect only accounted for 3.85% of the total variance, while the item-by-process effect accounted for 20.78% of the total variance. This finding indicated that while an item might receive a high rating on one process and a low rating on another process, the average ratings received by the items across different processes were comparable.

Rating Results

The final set of consensus ratings that was generated in the discussion for the reading items included in Forms E and F is displayed in Table 3. The final set of consensus item ratings revealed several major points. First, four components of the initial cognitive model — T4 (Using topical knowledge), TW1 (Using clues in other items), TW2 (Guessing), and TW3 (Using surface features of answer choices) — were considered by the raters as not really required for correctly answering the reading items included in both forms. Second, no additional processes were identified by the raters. The raters agreed that the set of processing components defined in the initial cognitive model covered all of the crucial processes required for correctly answering the reading items on the two MELAB test forms. Third, the final ratings indicated that correctly answering an item was often associated with the processes of using knowledge of particular lexical items, using syntactic knowledge, drawing inferences, locating specific details in text, and evaluating alternative options. Lastly, the rater results suggested that correctly answering an item often involved multiple cognitive processes. For example, to correctly answer Item 1 in Form E, the processes of recognizing words and word meanings, using syntactic knowledge, and drawing inferences are critically required, and the processes of

Table 3

Final Set of Consensus Ratings

Item	R1[1]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3
E1	2	1	2	1	1	2	1	2	1	1	0	0	0	0
E2	2	1	2	0	0	1	1	3	2	3	0	0	0	0
E3	2	1	1	2	2	2	5	4	3	2	0	0	0	0
E4	1	1	1	0	0	1	4	2	1	1	0	0	0	0
E5	2	2	2	0	0	2	4	1	1	2	0	0	0	0
E6	2	1	0	0	0	1	1	2	2	2	0	0	0	0
E7	1	1	1	0	0	1	1	2	1	1	0	0	0	0
E8	2	1	1	1	2	2	2	4	3	2	0	0	0	0
E9	0	1	0	0	1	0	1	3	1	0	0	0	0	0
E10	0	1	0	0	0	2	5	4	2	2	1	0	0	0
E11	2	1	1	0	0	0	4	3	2	1	0	0	0	0
E12	1	2	2	0	0	1	2	2	2	2	0	0	0	0
E13	1	2	1	2	1	0	4	2	1	0	0	0	0	0
E14	2	2	2	2	1	1	3	1	1	1	0	0	0	0
E15	1	2	2	0	0	1	5	2	2	2	0	0	0	0
E16	2	2	2	1	1	1	4	3	2	2	0	0	0	0
E17	1	1	1	1	1	1	4	3	2	2	0	0	0	0
E18	0	0	2	1	0	1	4	2	2	1	0	0	0	0
E19	1	0	2	2	1	2	1	2	1	2	0	0	0	0
E20	1	0	2	0	1	1	4	1	1	2	0	0	0	0
F1	1	0	1	1	1	0	1	2	3	0	0	0	0	0
F2	2	1	2	1	1	1	1	1	1	0	0	0	0	0
F3	1	0	2	0	0	0	1	2	2	1	0	0	0	0
F4	1	1	1	0	0	1	4	3	3	3	0	0	0	0
F5	1	0	1	0	1	0	3	3	2	0	0	0	0	0
F6	1	0	1	2	2	2	1	4	3	2	0	0	0	0
F7	0	0	1	1	1	2	4	4	2	1	0	0	0	0
F8	2	1	2	2	1	1	5	2	2	1	0	0	0	1
F9	2	0	2	2	2	1	5	1	2	2	0	0	0	0
F10	2	2	2	1	1	1	4	3	2	1	0	0	0	0
F11	2	1	1	2	2	2	3	4	3	1	0	0	0	0
F12	1	2	2	1	1	2	4	1	2	3	0	0	0	0
F13	2	1	2	1	2	2	4	3	2	3	0	0	0	0
F14	2	1	2	2	2	2	4	2	1	3	0	0	0	0
F15	2	2	2	1	2	2	4	3	2	3	0	0	0	0
F16	1	1	1	1	2	1	3	3	3	0	0	0	0	0
F17	1	0	2	0	0	0	4	3	2	2	0	0	0	0
F18	2	2	2	1	1	2	2	2	2	2	0	0	0	0
F19	2	1	1	2	2	2	2	1	2	2	0	0	0	0
F20	2	1	2	0	1	1	1	3	1	3	0	0	0	0

Note. [1] Abbreviations used in the table: R1 = Word Recognition Required; R2 = Vocabulary Knowledge Required; R3 = Syntactic Knowledge Required; R4 = Knowledge of Discourse Structure Required; R5 = Synthesis Required; R6 = Inference Required; R7 = Purpose of Information; T1 = Location of Information; T2 = Type of Match; T3 = Number of Plausible Distractors; T4 = Topical Knowledge Required; TW1 = Item Cues; TW2 = Guess; TW3 = Surface Feature of Options.

understanding infrequently used and/or specialized vocabulary, using knowledge of text structures, synthesizing information presented across multiple places in the text, understanding authors' pragmatic or rhetorical purposes, locating specific details in text, matching question to text, and evaluating alternative choices to eliminate seemingly correct options are somewhat involved.

To compare the cognitive item features across the two test forms, the frequency of the final consensus ratings were summarized separately for Form E and Form F, and the results are presented in Table 4. The numbers 0-5 contained in the first column of the

Table 4

Summary of Consensus Item Ratings for Form E and Form F

Rating	R1[1]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3
Form E														
0	3	3	3	11	10	3	–	–	–	2	19	20	20	20
1	8	11	7	5	8	11	6	3	9	6	1	0	0	0
2	9	6	10	4	2	6	2	9	9	11	0[3]	–	–	–
3	–[2]	–	–	–	–	–	1	5	2	1	–	–	–	–
4	–	–	–	–	–	–	8	3	–	–	–	–	–	–
5	–	–	–	–	–	–	3	–	–	–	–	–	–	–
Form F														
0	1	7	0	5	3	4	–	–	–	4	20	20	20	19
1	8	9	8	9	9	7	5	4	3	5	0	0	0	1
2	11	4	12	6	8	9	2	5	12	5	0	–	–	–
3	–	–	–	–	–	–	3	8	5	6	–	–	–	–
4	–	–	–	–	–	–	8	3	–	–	–	–	–	–
5	–	–	–	–	–	–	2	–	–	–	–	–	–	–

Note. [1] Abbreviations used in the table: R1 = Word Recognition Required; R2 = Vocabulary Knowledge Required; R3 = Syntactic Knowledge Required; R4 = Knowledge of Discourse Structure Required; R5 = Synthesis Required; R6 = Inference Required; R7 = Purpose of Information; T1 = Location of Information; T2 = Type of Match; T3 = Number of Plausible Distractors; T4 = Topical Knowledge Required; TW1 = Item Cues; TW2 = Guess; TW3 = Surface Feature of Options.

[2] The symbol "–"means that the point(s) in the rating scale were not applicable for that cognitive variable.

3The "0"s in the table means that no items included in the test form were coded for that point.

table represent the full range of points included in the rating scales for all cognitive variables, and the numbers in the next 14 columns represent the frequency of items coded for each point of the scale. For example, for the first variable "Word Recognition Required" (R1), which was measured using a 3-point scale (0 = Word recognition is not required to successfully complete the item; 1 = Word recognition is somewhat involved, but not critical to the successful completion of the item; and 2 = Word recognition is critical to successful completion of the item), 3 out of the 20 items included in Form E were coded as 0, 8 items were coded as 1, and 9 items were coded as 2. As points 4 and 5 in the first column of the table are not applicable for the variable R1, a dash is used in the corresponding cells, meaning "Not Applicable".

As Table 4 shows, the distributions of the item ratings were sometimes similar and sometimes different across the two forms. For the variables R7 (Purpose of Information), T4 (Topical Knowledge Required), TW1 (Item Cues), TW2 (Guessing), and TW3 (Surface Feature of Options), the distributions of the item ratings were comparable across the two test forms. For the variables T4, TW1, TW2, and TW3, almost all items included in both forms were coded as 0. This finding indicated that the processes using topical knowledge, using cues from the other items sharing the common passage, guessing, and using surface features of answer choices were not considered by the raters as required for correctly answering the items in both forms.

However, for the variables R1 (Word Recognition Required), R3 (Syntactic Knowledge Required), R4 (Knowledge of Discourse Structure Required), R5 (Synthesis

Required), and R6 (Inference Required), more items in Form F were coded 2 (Critical) than items in Form E. This finding indicated that more items in Form F than items in Form E were perceived by the raters as requiring the cognitively complex and demanding processes to arrive at the correct responses. These processes were: identifying words with advanced phonological and/or orthographical knowledge or recognizing the meaning of an unknown word with few context cues, understanding infrequently used or complex sentence structure, grammar, punctuation, or parts of speech, understanding complex discourse structure, working across multiple places in text, and generating the answer based on a respondent's global understanding of the entire passage, capability of predicting the continuation of arguments or events, or ability to relate information in the passage to the real world.

Further, for the variable T1 (Location of Information), while an identical number of items (3) in the two forms were coded 4 (The requested information is located across the entire passage or beyond the passage), more items in Form F than items in Form E (8 *vs.*5) were coded 3 (The requested information is located in the earlier part of the passage). These ratings indicated that, from the raters' perspective, more items in Form F requested information that might not have been in a respondent's short-term memory. For the variable T2 (Type of Match), more items in Form F than items in Form E lacked lexical overlap between the question and the text. This indicated that the raters perceived that more items in Form F required rewording, reordering, and integrating the idea structure of the passage to infer the information requested to correctly answer an item. For the variable T3 (Number of Plausible Distractors), while fewer items in Form F than items in Form E were coded by the raters as having two plausible distractors (5 *vs.* 11), more items in Form F were coded as having three plausible distractors (6 *vs.* 1). This indicated that the raters perceived that examinees taking Form F might need to make finer discriminations among the options to identify the correct answer.

Finally, of the 14 variables, R2 (Vocabulary Knowledge Required) appeared to be the only one with fewer items in Form F than items in Form E coded 1 or 2. This indicated that the raters perceived that the Form F items had a lower demand for understanding infrequently used and/or specialized vocabulary than the Form E items. As the cognitive demands of the items included in the two test forms differ, as perceived by the raters, the item difficulty distributions were likely to be different across the two forms.

Implications for the Model Revision

The results from coding the MELAB reading items in terms of the cognitive processes required to correctly answer each item had three major implications for revising the initial cognitive model. First, the rating results provided evidence that 10 processing components in the proposed initial cognitive model were required for correctly answering the MELAB reading items. These processing components were: R1 (Word recognition), R2 (Using vocabulary knowledge), R3 (Using syntactic knowledge), R4 (Using knowledge of discourse structure), R5 (Synthesis), R6 (Drawing inferences), R7 (Using pragmatic knowledge), T1 (Locating specific details in text), T2 (Matching the information given in the question stem to the relevant information in the text), and T3

(Evaluating alternative choices). Second, the rating results did not support the initial cognitive model in that four components of the model were not considered by the raters as required to arrive at the correct responses. These four components were: T4 (Using topical knowledge), TW1 (Using clues in other items), TW2 (Guessing), and TW3 (Using surface features of answer choices). It appeared that these four elements should be removed from the initial model. Third, no additional processes were identified by the raters, which suggested that the model overall covered all the crucial processes required for solving the reading items included in both forms. Hence, it appeared that no further processes needed to be added to the initial cognitive model.

Nevertheless, caution needs to be exercised when revising the model based on raters' perspective of the cognitive processes required to correctly answer the items. As all raters were senior graduate students who had used English in Canadian university settings for years, the cognitive processes identified by the raters as being needed to correctly answer an item may differ from the cognitive processes actually used by the MELAB target examinees. In this sense, verbal reports from the students representing the MELAB target examinees on the cognitive processes they use to correctly answer the reading items may shed light on the actual processes underlying the MELAB reading item performance and provide part of data triangulation. It was for this reason that the initial model was not modified until checking the consistency between the rating and the verbal report data, which is discussed in the next chapter.

---

[1] Due to test security, Form E and Form F of the MELAB reading test provided by the English Language Institute of the University of Michigan are not presented.

CHAPTER 5: VERBAL REPORTS OF STUDENTS

This chapter outlines the method used and presents the results for the second stage of the model validation — verbal reports from Mandarin-speaking students, who represent one of the largest language groups of the MELAB candidates. The results from the verbal reports analysis are then compared to the results from the raters' analysis of the test items in the previous chapter, to further refine the initial cognitive model and to determine the final set of item features for use in the statistical model described in Chapter 6.

Method

Participants

The participants for Phase 2 were Mandarin-speaking students who started their undergraduate or graduate programs at the University of Alberta in the fall of 2005 or winter 2006 and who had resided in English-speaking countries for no more than nine months. These students were selected because they were considered (1) to be literate in their L1 as a result of at least 11 years of basic education in China, and (2) to have acquired advanced-level English reading proficiency, that is, have mastered the English language knowledge and processing strategies required for reading academic texts at the college level. The selection of nine months was based on three studies in which verbal reports of the L2 students were used to investigate the processes they used to answer reading test items. Anderson et al. (1991) limited their participants to those who had studied in the US from nine weeks to nine months; Abbott (2005) limited her participants to those who had resided in Canada for less than two years; and Cohen and Upton (2006) used undergraduates and graduates who had resided in the US from nine to 84 months.

A recruiting letter (see Appendix F) was sent via the Chinese Students and Scholars Association mailing list, which is available to all Chinese students at the university, 10 days prior to the data collection. The recruiting letter described the nature of the investigation, explained what would be expected of the participants, clarified the participant selection criteria, and asked for their voluntary participation. The students were assured that their names would not be used in any published work, and that their verbal reports would not be used for any purposes except by the researcher to determine how they answered the test items. The students who replied and expressed an interest in participation were contacted by the researcher. The sample size was set at 18 (2 for the pilot study and 16 for the main study). The recruited participants were randomly assigned to take either Form E or Form F, with equal numbers of participants for each form. This sample size was considered reasonable, as prior research had shown that all properties of the categories could be identified and no new relevant data emerged from the verbal reports after analyzing eight to ten protocols (Abbott, 2005; Wu, 1998).

Materials

Background questionnaire. A background questionnaire was used to interview all participants to obtain information on their age, gender, level of education, discipline of study, time spent studying English in their home country, and length of residence in English-speaking countries (see Appendix G). The use of the background questionnaire had three main purposes: (1) to determine the participants' demographic characteristics,

(2) to establish researcher-participant rapport, and (3) to make sure all participants met the selection criteria described earlier. To ensure that the participants understood the questions, the interviews were conducted in the participants' first language (Chinese). The participants' verbal responses to the background questions and test items were recorded using a digital audio recorder.

Verbal report practice tasks. Verbal report practice tasks were used to familiarize the participants with the verbal report method. In order to avoid misleading the participants, reading items were not used as practice tasks. Instead, the three verbal report tasks used by Ericsson and Simon (1993) to illustrate the concurrent and retrospective verbal report method were employed (see Appendix H).

MELAB test forms. Form E and Form F of the MELAB reading test used in the first stage of the model validation continued to be used for this investigation (see Chapter 4 for the description of the two test forms).

Data Collection

Pilot study. A small-scale pilot study was conducted to refine the data collection procedure and the protocol analysis procedure to be used to analyze the responses of the students in the main study. Two Chinese students enrolled in a graduate program at the University of Alberta in the fall of 2005 participated in the pilot study, with one taking Form E and the other taking Form F. For each student, one session of approximately two hours was scheduled to complete the consent form, background questions, training on giving verbal reports, 20-item reading test, and concurrent and retrospective verbal reports. The participants could respond in their L1 (Mandarin), L2 (English), or both. The data collected were transcribed, translated into English, and coded by the researcher.

It was determined that the data collection and analysis procedures used in the pilot study needed to be revised. First, based on the observation that one session to complete all the tasks described above extended the workload beyond what could reasonably be tolerated by the participants, the data collection period was split into two sessions, with two reading passages and their associated items administered in each session.

Second, the retrospective verbal reports conducted immediately following the completion of the concurrent verbal reports for each test item resulted in sparse concurrent data. In other words, the students kept silent while answering the items, and verbalized their thoughts only after they had completed the items. Therefore, students were instructed to give a verbal report on each item as it was completed, and retrospective verbal reports were conducted immediately following the completion of the set of five items for each passage.

Third, the type and frequency of probes distracted, and, at times, misled the participants. Hence, researcher probes were replaced by non-mediated verbal probes successfully used by Abbott (2005).

Fourth, the verbal report data, which were collected entirely in Mandarin, were too extensive to transcribe, translate, and code for analysis in a timely manner. To improve the efficiency and economy of data analysis and be more faithful to the original data, the students' verbal reports were not translated into English, except for those portions presented in this chapter to illustrate the students' thinking.

Lastly, the verbal data had been intended to validate both the components of the initial cognitive model and the cognitively-based item features coded by the raters. However, it was found that the verbal data coded using the coding scheme based on the components of the initial cognitive model could not be completely matched to the rating data. Therefore, two-step intensive analyses of the verbal data were conducted. That is, the verbal data were first classified using the coding scheme based on the cognitive processes defined in the initial cognitive model (i.e., the first two columns of Table 1 in Chapter 3). Once this step was done, the verbal report data were further coded using the finer coding scheme based on the rating scales listed in the last column of Table 1. For example, using the initial coding scheme based on the components of the initial cognitive model, the verbal data for a given item were classified into two categories, drawing inferences and locating specific details in the text. Then, using the finer coding scheme based on the rating scales, the verbal data were further coded as drawing low-level inferences or high-level inferences and in which part of the text the requested information could be found (i.e., 1st section, 2nd section, 3rd section, or the entire passage).

Main study. The verbal report data were collected in individual sessions at the University of Alberta during the winter 2006. The respondents were told that they could respond in Mandarin, English, or both. As pointed out above, data from each participant were collected during two separate sessions scheduled on two different days within a week, with the first two passages and their associated items administered on day 1 and the remaining two passages and their associated items administered on day 2. To maximize the consistency among the sessions, standardized procedures and instructions were followed for each session (see Appendix G).

On day 1, the researcher met with each participant in an office at the university. The researcher and participant sat side-by-side at a table on which there was a microphone, a digital audio recorder, and a folder containing the experimental materials. These materials included a consent form (see Appendix I), a sheet of instructions and background questions, three practice tasks, and either Form E or Form F of the MELAB reading test. Before beginning the formal data collection, the researcher explained the nature of the task and the procedures to be followed. After the participant read and signed the consent form, the researcher interviewed the participant about his/her age, education level, major, length of time studying English, and length of residence in English-speaking countries.

Considering that the participants might not be familiar with concurrent and retrospective verbal report methods, the researcher then provided instructions and the three tasks for them to practice their verbal reporting skills. Once the participants were accustomed to the verbal report procedures and had no questions, the digital audio recorder was turned on and the first two passages with their accompanying 10 items from the assigned form were administered. Participants were asked to read the passages and answer the accompanying items as if they were taking a real reading test. They were asked to verbally express their thought processes while answering the items (concurrent reports), and upon completing the set of five items for each passage, to describe aloud their remembrance about the thought processes they used to answer the item (retrospective reports). The participants were instructed to verbalize whatever was on their minds while and after they completed each item and to talk constantly. If the participants remained silent for longer than 10 seconds, they were prompted by the researcher to keep talking.

After completing the verbal report for both passages, the participants were asked by the researcher to rate their familiarity with the passage topics using the scale 0 = familiar, 1= somewhat familiar, 2 = not familiar. The topic familiarity ratings were intended to determine whether or not topical knowledge was a source of processing difficulty that affected the difficulty of the MELAB reading items, as indicated in the initial cognitive model. On day 2, the participants were asked to complete the remaining two passages with their accompanying 10 items on the test form. The procedures followed on day 2 were the same as the procedures used on day 1.

Verbal Protocol Procedures

Coding the verbal report data. The students' answers were marked to the key and their verbal reports were transcribed verbatim and typed into computer by the researcher. The researcher reviewed each of the verbal reports and coded them for the cognitive processes used for both correct and incorrect responses. The concurrent and retrospective data were combined for each item for analysis. The cognitive processes defined in the initial cognitive model were used as a coding scheme for classifying the verbal report data. The statement or phrase of the verbal reports associated with each cognitive process was segmented and assigned a code. Additional processes gleaned from the transcripts were categorized and added to the existing list of cognitive processes.

Upon completion, the verbal report data were recoded in a finely-tuned manner using the detailed coding scheme based on the rating scales listed in the last column of Table 1. As the coding proceeded, it became clear that the verbal data were not sensitive enough to make finely-tuned distinctions between the categories "Somewhat involved in correctly answering the item" and "Critical for correctly answering the item" for the processes R1 (Word recognition), R2 (Using vocabulary knowledge), R3 (Using syntactic knowledge), R4 (Using knowledge of discourse structure), and T4 (Using topical knowledge). Therefore, these two categories were collapsed. That is, the verbal data regarding these five processing categories were simply coded as 0 if they were not reported and coded 1 if reported. All the individual occurrences of the cognitive processes were identified and coded, and the total number of times each process was verbalized for the correct responses was counted.

Coding reliability. To evaluate the reliability of assigning the processes used by the students to the various processing categories in the initial model, an independent rater (i.e., a colleague of the researcher, who had comparable expertise as the researcher and no experience with the study) coded 37.5% of the verbal report data first using the initial coding scheme and then using the finer coding scheme mentioned above. To ensure the confidentiality of the data and research information, the independent rater was asked to read and sign a confidentiality agreement (see Appendix J). Prior to data coding, the independent rater was trained in the data coding procedures. During the training, the researcher discussed the coding schemes with the rater, demonstrated the coding, and provided the rater with a chance to practice using the verbal reports from one of the participants. After the training, the rater independently coded six verbal reports, three of which were randomly selected from the Form E participants and three randomly selected from the Form F participants. The codes assigned by the independent rater were then compared to those assigned by the researcher. The reliability of assigning processes to the various processing categories was evaluated. Consistency was defined as the extent to which the verbal report data segments were coded using the same processing categories by both raters. The percentage of total agreement between the researcher and the independent rater was 82.8%, which indicated that the cognitive processes segments were consistently coded.

Supplementary data analysis. The participants' responses to the background questions and topic familiarity ratings were entered into the SPSS 13.0 and verified for 100% accuracy. Descriptive statistics were calculated to determine the participants' demographic characteristics and familiarity with the MELAB reading topics.

Comparing verbal reports to ratings. To determine the final set of item features, the cognitively-based item features coded by the raters were compared to the frequency of the cognitive processes inferred from the students' verbal reports for the correct responses. Because of the variety in the total number of correct responses per item, a simple count of process frequencies would have distorted the importance of any given process. Hence, the raw frequencies of the processes used by the students for the correct responses to each item were converted into ratio scores, that is, the number of students reporting the use of a given process for the correct responses to a given item in relation to the total number of students who correctly answered that item. Then, criteria were developed to rigorously distinguish levels of consistency between the rating data and the verbal report data. The cutoff points were used as follows: The two data sources were considered as highly consistent when the ratio scores > 0.5, which meant a majority of the students who correctly answered a given item reported the use of a given process coded by the raters; moderately consistent when the ratio scores >0.25 and ≤ 0.5; low consistent when the ratio scores ≤ 0.25 and >0; and not consistent when the ratio scores = 0. Based on these criteria, consistency between the two data sources were checked, and reasons for the contradictory findings were examined. Based on the results of comparison, the processing components of the initial cognitive model and cognitive variables used for rating were refined as necessary. Given the complexity of validating the cognitive model based on frequency of the verbal data and comparison of the verbal data to the rating data, the components of the cognitive model derived from these analyses were analyzed in a partly empirical and partly intuitive way, relying more on qualitative criteria than on any hard and fast quantitative measures.

Results

Biographic Description

Table 5 presents the participants' code (e.g., E1 means Form E Participant 1), age, gender, level of education, discipline of study, time spent studying English in their home country (TOE), and length of residence in English-speaking countries (LOR), as determined by the background questions answered by each participant at the beginning of session 1. The last column of Table 5 shows the participants' total scores on the form they responded to.

Table 5

Characteristics of the Verbal Report Participants

Code	Age	Gender	Education	Discipline	TOE[1]	LOR[2]	Score
E1	24	F	Undergraduate	Mathematics	9	2	13/20
E2	29	M	Graduate	Material Science	12	5	18/20
E3	32	M	Graduate	Psychology	13	5	18/20
E4	19	F	Undergraduate	Biology	11	6	15/20
E5	28	M	Graduate	Linguistics	13	6	19/20
E6	25	M	Graduate	Electrical Engineering	10	1	16/20
E7	24	F	Graduate	Education	8	6	18/20
E8	32	F	Graduate	English Literature	19	6	16/20
F1	31	F	Graduate	Civil Engineering	8	9	11/20
F2	33	M	Graduate	Computer Science	13	5	14/20
F3	27	M	Graduate	Marketing	13	6	19/20
F4	25	M	Graduate	Chemical Engineering	8	7	18/20
F5	26	M	Graduate	Pharmacy	13	6	19/20
F6	22	F	Graduate	Statistics	10	6	19/20
F7	30	M	Graduate	Computer Science	13	1	16/20
F8	26	F	Graduate	Education	13	6	15/20

Note. [1]TOE = Time spent studying English in their home country (years)
[2]LOR = Length of residence in English speaking countries (months)

As shown in Table 5, the interview sample included 2 undergraduates (i.e., E1 and E4) and 14 graduates. Eight students (4 males and 4 females, of whom two were undergraduates) responded to Form E, and eight students (5 males and 3 females) responded to Form F. The participants were from a variety of disciplines. At the time when they participated in the study, they had spent at least eight years studying English in their home country and resided in English-speaking countries for no more than nine months. The participants had a minimum score of 11 out of 20, which indicated that they were at a proficiency level where they had acquired at least basic understanding of the test passages and would be able to provide accurate reports of the processes they used to answer reading test items rather than simply guessing.

The mean, standard deviation, and range of the participants' age, TOE, LOR, and total scores were calculated separately for Form E and Form F. The results are displayed in Table 6. As shown in the table, the mean age of the participants for Form E was 26.62 (*SD* = 4.47), and for form F was 27.50 (*SD* = 3.59). The mean length of time spent in studying English in their home country (TOE) was 11 years and 10 months for the Form E participants (*Mean* = 11.87, *SD* = 3.40), and 11 years and 5 months for the Form F participants (*Mean* = 11.38, *SD* = 2.33). At the time of completing the verbal reports, the Form E participants had resided in English-speaking countries for 4.62 months on average (*SD* = 2.00), and the Form F participants had resided in English-speaking countries for 5.75 months on average (*SD* = 2.25). The mean score of the participants for Form E was 16.63 (83.2%) (*SD* = 2.00), and for Form F was 16.38 (81.9%) (*SD* = 2.92). The mean differences between the Form E and Form F participants in each of the four variables were tested using independent samples *t*-tests. The results indicated that there was no significant difference between the two samples at the 0.05 level of significance

($t_{age} = -0.432, p > 0.05; t_{TOE} = 0.343, p > 0.05; t_{LOR} = -1.058, p > 0.05; t_{score} = 0.200, p > 0.05$).

Table 6

Descriptive Statistics for the Verbal Report Participants

Variables	Form E			Form F		
	Mean	*SD*	*Range*	*Mean*	*SD*	*Range*
Age	26.62	4.47	13.00	27.50	3.59	11.00
TOE[1]	11.87	3.40	11.00	11.38	2.33	5.00
LOR[2]	4.62	2.00	5.00	5.75	2.25	8.00
Score	16.63	2.00	6.00	16.38	2.92	8.00

Note: [1]TOE = Time spent studying English in their home country (years)
[2]LOR = Length of residence in English speaking countries (months)

Topic Familiarity

Table 7 displays the distributions of the participants' ratings on their familiarity with the topics of the reading passages included in Forms E and F. The passages in Form E were on the topics of sea life (P1), human mate selection (P2), rock formation (P3), and bread (P4), and in Form F were on the topics of pilgrimage (P1), sleep (P2), agriculture (P3), and fleas (P4). As shown in the table, with the exception of the first and third passages in Form E, a majority of the participants indicated that they were either familiar or somewhat familiar with the passage topics. The third passage in Form E, which was about rock formation, appeared to be the least familiar to the participants, with 2 participants indicating familiarity and 6 indicating unfamiliarity. The first passage in Form E, which was about sea life, was the second least familiar to the participants, with 4 participants indicating familiar or somewhat familiar and 4 indicating unfamiliar. The next two passages with relatively less familiar topics were the last passage in Form E (about bread) and the last passage in Form F (about flea), respectively. The second passage in Form F, which was about sleep, appeared to be the most familiar to the participants, with 5 participants indicating familiar, 2 indicating somewhat familiar, and only 1 indicating unfamiliar.

Table 7

Distributions of MELAB Topic Familiarity by Test Form

Rating Scale	Form E (*N* = 8)				Form F (*N* = 8)			
	P1[1]	P2	P3	P4	P1	P2	P3	P4
Familiar	1	3	2	0	3	5	3	1
Somewhat Familiar	3	4	0	5	4	2	4	4
Not Familiar	4	1	6	3	1	1	1	3

Note.[1]P represents passage. For example, P1 means Passage One.

Comparison of the topic familiarity distributions between Form E and Form F revealed that Form E had more students who lacked topic familiarity than Form F. Given that no significant differences were found between the Form E and Form F students' mean scores at the 0.05 level of significance, it appeared that topic familiarity was not that important for these two forms. This finding was not surprising, considering that the MELAB test specifications specify that topic knowledge is not required to understand the passages or to answer the items and that any bias toward examinees of a particular educational background should be avoided when selecting the test passages.

Nevertheless, considering that the topics of the MELAB test passages are intended to be accessible to all target examinees and not to be easier or more difficult for examinees of any particular field of study or from any particular country, cultural, or experiential background (ELI, 2003), it was surprising that some participants indicated that they were not familiar with the passage topics. An examination of their verbal reports provides some explanations of why they reported unfamiliarity. The following are examples of the participants' comments on topic unfamiliarity. The notes in the bracket mean the test form, passage, and student. For example, EP2S1 means Form E Passage 2 Student 1.

a) "I was not familiar with this passage topic, especially the content of the passage. Indeed, movement of the earth's crust was far from my background knowledge as a liberal art student. I had to read the passage carefully and then solved the items based on what was said in the passage. I felt I could answer the items, but could not tell exactly what was going on in this passage. I just didn't know what it (the passage) talked about. For example, I didn't understand the proper nouns and there was no way to guess their meanings." [EP3S8]

b) "Judging from my common sense, I was not familiar with this topic. I met such a topic only when I took a test. I definitely lacked the knowledge about rock formation. I didn't know much about rock formation, especially when I had no idea about this key word [magma]. Fail to understand this word affected my understanding about the passage content. I think understanding the meaning of key words in the passage is the key to understanding the passage content." [EP3S5]

c) "I said I was not familiar with the topic of the first passage. In fact, my unfamiliarity was mainly because my unfamiliarity with the vocabulary. In retrospect, actually I was somewhat familiar with all the passage topics. All test passages are about popular science and the texts are completely understandable." [EP1S1]

d) "I was not familiar with this topic at all. A lot of words in the passage were not familiar to me. I had to guess their meanings. Besides, there were so many proper nouns and I did not know how to pronounce them. Although proper nouns did not affect understanding too much, they did affect understanding and recall to some degree. A big problem is that I feel so frustrated, so annoyed, and so impatient whenever reading texts with lots of unfamiliar words. I just have no interest reading it at all." [EP1S3]

e) "Even if I am quite familiar with bread and eat it everyday, I doubt I am familiar with the content of this passage. I didn't know the mechanisms for producing bread and the logical explanations for producing different types of breads. My common knowledge didn't tell me the relationships between producing different breads and economies of a country and the author's and economists' opinions. It is like when you read a paper in your own field. Even if you are familiar with its topic, you don't know its content, because different people may have different ideas and get different conclusions. You never know what they are talking about until you read their stuff. I have to read the passage anyway. Hence, I think 'topic familiarity' is not an important issue to purse, and indeed, I feel topic familiarity is a very ambiguous term" [EP4S6]

f) "I was not familiar with the topic of this passage. Theoretically, it is understandable that the topics of the test passages are not familiar or somewhat familiar, since the passages with very familiar topics are normally not used for test purpose. Say, this passage, I definitely knew flea. I even knew flea was the insect that jumped the highest in relative to its height. However, I didn't know its physical characteristics, work mechanisms, and evolution. I knew about the plague endemic in the rat population, which caused death of tens of thousands of people during the 14th century. But I didn't know how the plague found its way to the human population. I didn't know the rats got the plague from the fleas. I had thought it was caused by the food taken by rats."[FP4S4]

The students' comments revealed three major reasons for why some of them reported unfamiliarity with certain passage topics. First, as can be seen from the comments given by the students EP3S8, EP3S5, EP1S1, and EP1S3, the appearance of proper nouns and failure to understand the meaning of key vocabulary in the passage affected their interest in reading the passages and understanding about the passage content. It appeared that the unfamiliarity with the passage topics reported by these students was likely due to their unfamiliarity with the infrequently used and/or specialized vocabulary. Second, as indicated by the students EP3S8 and EP3S5, who majored in English Literature and Linguistics, respectively, certain passage topics like rock formation might have bias toward the students from certain academic backgrounds. Third, as indicated by the students EP4S6 and FP4S4, topic familiarity appeared to be a vague term, as clear-cut descriptions to distinguish familiarity from unfamiliarity or different levels of familiarity were lacking. Consequently, evaluation of topic familiarity became very subjective in nature. For instance, certain students reported familiarity with the topic about bread based on the criteria that they were familiar with the bread. But some students reported unfamiliarity with this topic based on their criteria that they did not know about the details described in the passage, such as the mechanisms for producing the bread, though they knew about the bread. Maybe it is true to say "topic familiarity is not an important issue to pursue", as commented by student EP4S6.

Validating the Components of the Initial Model with the Verbal Reports

While both correct and incorrect responses were coded, only the cognitive processes underlying the correct responses are presented, because the interpretation of the correct responses to items was considered most pertinent to the validity of a test (Alderson, 2000) and the development of a cognitive processing model (VanderVeen, 2004). Tables 8 and 9 present the frequency of each processing component of the initial cognitive model used by the students who correctly answered each item included in Form E and Form F, respectively. In both tables, the first column displays the test form and item number (e.g., E1 means Form E Item 1). The second column summarizes the number of correct responses for each item (e.g., 7 means 7 out of 8 participants correctly answered that item). The next 15 columns display the frequency of each cognitive process used by one or more students who correctly answered the item.

As shown in Tables 8 and 9, the verbal reports for a total of 134 correctly answered items in Form E and 131 correctly answered items in Form F were analyzed. The

Table 8

Frequencies of Cognitive Processes Used for the Correct Responses (Form E)

Item	#✓[1]	R1[2]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3	T5[3]
E1	4	2	1	2	0	3	3	0	3	2	3	0	0	0	0	1
E2	7	4	0	2	0	1	0	0	7	1	4	0	0	0	0	3
E3	5	0	0	0	3	1	0	4	0	0	5	0	0	0	1	1
E4	7	0	0	3	0	0	6	0	7	2	4	0	0	0	0	2
E5	3	0	0	1	0	0	3	0	3	2	0	0	0	0	0	3
E6	7	7	0	0	2	0	0	0	7	0	3	0	0	0	0	1
E7	8	0	0	0	0	0	0	0	8	0	2	0	0	0	0	4
E8	7	0	0	0	1	1	3	0	5	0	5	1	0	3	0	2
E9	8	0	0	1	0	2	0	0	8	3	6	0	0	0	0	4
E10	5	0	0	0	2	0	2	2	0	0	5	2	0	2	0	2
E11	8	0	0	0	0	0	0	0	8	2	5	0	0	0	0	4
E12	8	1	0	1	0	0	4	0	8	3	4	0	0	1	0	4
E13	8	6	0	0	7	0	0	2	8	5	2	0	0	0	0	5
E14	8	2	0	0	1	1	6	0	8	3	6	0	0	2	0	4
E15	7	1	0	3	1	1	5	1	7	4	3	0	0	1	0	4
E16	7	7	0	0	1	0	4	0	6	2	4	0	0	1	0	1
E17	5	2	0	0	3	4	3	1	5	2	5	0	0	0	0	1
E18	6	0	0	5	0	0	4	1	6	2	4	0	0	0	0	2
E19	8	1	0	3	2	1	7	0	7	4	6	0	0	1	0	3
E20	8	0	0	5	3	2	5	2	8	3	6	0	0	0	0	3
Total	134	33	1	26	26	17	55	13	119	40	82	3	0	11	1	54

Note. [1] #✓ means the number of correct responses.

[2] Abbreviations used in the table: R1 = Using word recognition; R2 = Using vocabulary knowledge; R3 = Using syntactic knowledge; R4 = Using knowledge of discourse structure; R5 = Synthesis; R6 = Drawing inferences; R7 = Recognizing authors' pragmatic or rhetorical purposes; T1 = Locating specific details in text; T2 = Matching question to text; T3 = Evaluating alternative options; T4 = Using topical knowledge; TW1 = Using item cues; TW2 = Guessing; TW3 = Using surface features of answer choices.

[3] T5 = Additional processing category emerged from the data– metacognitive strategies.

Table 9

Frequencies of Cognitive Processes Used for the Correct Responses (Form F)

Item	#✓[1]	R1[2]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3	T5[3]
F1	8	3	0	3	2	3	0	7	6	2	6	0	0	0	0	3
F2	7	1	0	4	2	0	5	0	7	3	5	0	0	0	0	5
F3	8	4	0	7	0	0	3	1	8	2	3	0	0	0	0	4
F4	6	4	0	0	0	0	5	0	5	2	3	0	0	1	0	2
F5	8	1	0	3	0	0	0	3	8	3	5	0	0	0	0	4
F6	7	2	0	0	2	4	6	2	2	2	6	0	0	2	0	5
F7	8	0	0	0	2	3	8	1	7	2	6	2	0	1	0	5
F8	7	0	0	0	1	3	5	0	5	3	2	0	0	0	0	4
F9	7	0	0	0	1	1	0	0	5	2	4	0	0	0	0	2
F10	7	2	0	1	0	0	0	0	7	2	5	1	0	0	0	4
F11	7	0	0	0	4	4	0	4	0	0	6	0	0	0	0	4
F12	6	3	0	3	1	0	4	1	6	3	3	1	0	0	0	3
F13	6	1	0	0	2	1	5	1	5	2	4	1	0	1	0	3
F14	5	4	0	5	0	0	0	0	5	4	3	2	0	1	0	3
F15	4	0	0	1	0	1	0	2	4	4	3	1	0	0	0	1
F16	8	0	0	0	7	8	0	5	0	0	6	0	0	0	0	4
F17	4	0	0	3	0	1	0	0	4	2	2	1	0	0	0	2
F18	6	2	0	4	1	0	3	0	6	1	5	1	0	2	0	5
F19	5	2	0	4	0	2	0	1	3	1	3	1	0	1	0	2
F20	7	5	0	4	0	0	3	0	7	1	5	2	0	1	0	3
Total	131	34	0	42	25	31	47	28	100	41	85	13	0	10	0	68

Note. [1] #✓means the number of correct responses.

[2] Abbreviations used in the table: R1 = Using word recognition; R2 = Using vocabulary knowledge; R3 = Using syntactic knowledge; R4 = Using knowledge of discourse structure; R5 = Synthesis; R6 = Drawing inferences; R7 = Recognizing authors' pragmatic or rhetorical purposes; T1 = Locating specific details in text; T2 = Matching question to text; T3 = Evaluating alternative choices; T4 = Using topical knowledge; TW1 = Using item cues; TW2 = Guessing; TW3 = Using surface features of answer choices.

[3] T5 = Additional processing category emerged from the data– metacognitive strategies.

cognitive processes used by the participants to correctly answer the reading items in both forms covered 11 (out of 14) components of the initial cognitive model. The three processing components not supported by the students' verbal report data were R2 (Using vocabulary knowledge), TW1 (Using item cues), and TW3 (Using surface features of answer choices). This finding was not surprising, considering that the students were all advanced-level ESL students who have mastered basic English language knowledge and processing strategies required for reading university-level academic texts, and that using item cues and surface features of answer choices were irrelevant to the constructs assessed by the MELAB reading.

Additional processes emerged from the students' verbal reports included assessing the correctness of the response to the test item, assessing the characteristics of the test item to determine the feasibility of correctly answering it, assessing what is required to complete the item, planning the order of steps to be taken to complete the items, and making verifications in order to answer the item correctly. Based on the literature (e.g., Alderson, 2000; Bachman & Palmer, 1996; Phakiti, 2003), these processes were classified into a new processing category of test management called the metacognitive strategies (T5).

The marginal frequency of each cognitive process used by the Form E and Form F students to arrive at the correct responses (i.e., the last rows of Tables 8 and 9) were compared, and the results are displayed in Figure 6. In the figure, the Y-axis represents the frequency of the cognitive processes used by the students, and the X-axis displays the 15 cognitive processing categories inferred from the students' verbal reports. Using Form E as the base form, these cognitive processes are ordered in terms of the descending frequency of Form E. For all processes, the frequencies of the processes used for Forms E and F are presented as pairs, with Form E being the left bar and Form E the right.

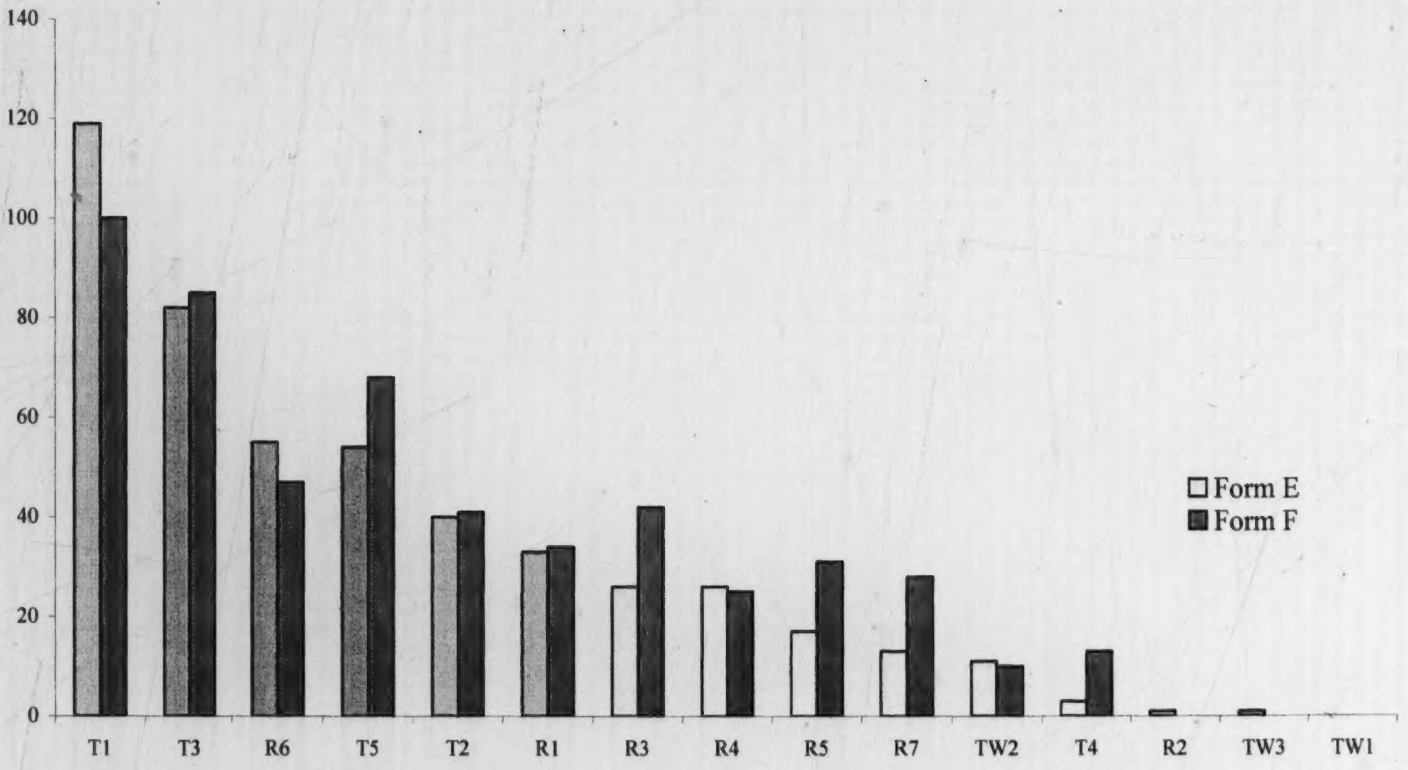


Figure 6. Frequency of the cognitive processes used by the Forms E and F students[1].

Note. [1]Abbreviations used in the table: R1 = Word recognition; R2 = Using vocabulary knowledge; R3 = Using syntactic knowledge; R4 = Using knowledge of discourse structure; R5 = Synthesis; R6 = Drawing inferences; R7 = Using pragmatic knowledge; T1 = Locating specific details in text; T2 = Matching question to text; T3 = Evaluating plausible distractors; T4 = Using topical knowledge; TW1 = Using item cues; TW2 = Guessing; TW3 = Using surface feature of options; T5 = Metacognitive strategies.

Two features are immediately apparent in the figure. First, a majority of the cognitive processes have comparable frequencies of occurrences between Forms E and F, and only a small number of cognitive processes have different frequencies of occurrences between the two forms. The cognitive processes having comparable frequencies of occurrences between the two forms include T3 (82 times for Form E *vs.* 85 times for Form F), T2 (40 times for Form E *vs.* 41 times for Form F), R1 (33 times for Form E *vs.* 34 times for Form F), R4 (26 times for Form E *vs.* 25 times for Form F), TW2 (11 times for Form E *vs.* 10 times for Form F), R2 (once for Form E *vs.* zero for Form F), TW3 (once for Form E *vs.* zero for Form F), and TW1 (zero for both forms). The cognitive processes having different frequencies of occurrences between the two forms include T1 (119 times for Form E *vs.* 100 times for Form F), R6 (55 times for Form E *vs.* 47 times for Form F), T5 (54 times for Form E *vs.* 68 times for Form F), R3 (26 times for Form E *vs.* 42 times for Form F), R5 (17 times for Form E *vs.* 31 times for Form F), R7 (13 times for Form E *vs.* 28 times for Form F), and T4 (3 times for Form E *vs.* 13 times for Form F).

Second, as can be seen in the figure, the rank orderings of the cognitive processes in terms of their frequencies of occurrences display similar patterns for both forms. For example, for both forms, T1 (Locating specific details in text) and T3 (Evaluating alternative choices) were the two most frequently reported cognitive processes for the correct responses. The Form E students reported the use of T1 for the correct responses for a total of 119 times and T3 for a total of 82 times. The Form F students reported the use of T1 for the correct responses for a total of 100 times and T3 for a total of 85 times.

Moreover, for both forms, R2 (Using vocabulary knowledge), TW1 (Using clues in other items), and TW3 (Using surface features of answer choices) were not really reported by the students for the correct responses. For the remaining 10 cognitive processes, only minor differences were discovered in their rank orders between the two forms. Specifically, these cognitive processes reported by the Form E students for the correct responses, in descending order, were R6, T5, T2, R1, R3, R4, R5, R7, TW2, and T4, and for Form F, in descending order, were T5, R6, R3, T2, R1, R5, R7, R4, T4, and TW2.

Comparing the Item Feature Ratings to the Verbal Reports

Tables 10 and 11 present the results of comparisons between the final set of consensus ratings presented in Table 4 and the frequency of the cognitive processes inferred from the students' recount for the correct responses and coded using the finer coding scheme based on the rating scale. The table was again set up as an item by process matrix, with rows containing the items and columns containing the 14 cognitive elements. Each cell of the table contained two numbers: the left one is the item feature coded by the raters and the right one in the bracket shows the frequency ratio of the corresponding cognitive process used by the students to correctly answer the given item. As mentioned earlier, frequency ratio was defined as the number of students reporting the use of the given process to reach the correct responses to the given item divided by the number of students who correctly answered that item. For example, the number 0/7 in a bracket means that 7 students correctly answered that item, and that the rating on a given process for that item could not be supported by the verbal reports from any of these 7 students.

Table 10

Comparisons between Verbal Data and Rating Data (Form E)

Item	R1[1]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3
E1	2(2/4)	1(1/4)	2(2/4)	1(0/4)	1(3/4)	2(2/4)	1(0/4)	2(3/4)	1(2/4)	1(2/4)	0(4/4)	0(4/4)	0(4/4)	0(4/4)
E2	2(4/7)	1(0/7)	2(2/7)	0(7/7)	0(6/7)	1(0/7)	1(0/7)	3(7/7)	2(1/7)	3(4/7)	0(7/7)	0(7/7)	0(7/7)	0(7/7)
E3	2(0/5)	1(0/5)	1(0/5)	2(3/5)	2(1/5)	2(0/5)	5(4/5)	4(5/5)	3(5/5)	2(3/5)	0(5/5)	0(5/5)	0(5/5)	0(4/5)
E4	1(0/7)	1(0/7)	1(3/7)	0(7/7)	0(7/7)	1(6/7)	4(0/7)	2(7/7)	1(2/7)	1(2/7)	0(7/7)	0(7/7)	0(7/7)	0(7/7)
E5	2(0/3)	2(0/3)	2(1/3)	0(3/3)	0(3/3)	2(2/3)	4(0/3)	1(3/3)	1(2/3)	2(0/3)	0(3/3)	0(3/3)	0(3/3)	0(3/3)
E6	2(7/7)	1(0/7)	0(7/7)	0(5/7)	0(7/7)	1(0/7)	1(0/7)	2(7/7)	2(0/7)	2(3/7)	0(7/7)	0(7/7)	0(7/7)	0(7/7)
E7	1(0/8)	1(0/8)	1(0/8)	0(8/8)	0(8/8)	1(0/8)	1(0/8)	2(8/8)	1(0/8)	1(2/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)
E8	2(0/7)	1(0/7)	1(0/7)	1(1/7)	2(1/7)	2(2/7)	2(0/7)	4(5/7)	3(0/7)	2(3/7)	0(6/7)	0(7/7)	0(4/7)	0(7/7)
E9	0(8/8)	1(0/8)	0(7/8)	0(8/8)	1(2/8)	0(8/8)	1(0/8)	3(8/8)	1(3/8)	0(2/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)
E10	0(5/5)	1(0/5)	0(5/5)	0(3/5)	0(5/5)	2(2/5)	5(2/5)	4(3/5)	2(0/5)	2(4/5)	1(3/5)	0(5/5)	0(3/5)	0(5/5)
E11	2(0/8)	1(0/8)	1(0/8)	0(8/8)	0(8/8)	0(8/8)	4(0/8)	3(8/8)	2(2/8)	1(5/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)
E12	1(1/8)	2(0/8)	2(1/8)	0(8/8)	0(8/8)	1(4/8)	2(0/8)	2(8/8)	2(3/8)	2(4/8)	0(8/8)	0(8/8)	0(7/8)	0(8/8)
E13	1(6/8)	2(0/8)	1(0/8)	2(7/8)	1(0/8)	0(8/8)	4(2/8)	2(8/8)	1(5/8)	0(6/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)
E14	2(2/8)	2(0/8)	2(0/8)	2(1/8)	1(1/8)	1(6/8)	3(0/8)	1(8/8)	1(3/8)	1(3/8)	0(8/8)	0(8/8)	0(6/8)	0(8/8)
E15	1(1/7)	2(0/7)	2(3/7)	0(6/7)	0(6/7)	1(5/7)	5(1/7)	2(7/7)	2(4/7)	2(3/7)	0(7/7)	0(7/7)	0(6/7)	0(7/7)
E16	2(7/7)	2(0/7)	2(0/7)	1(1/7)	1(0/7)	1(4/7)	4(0/7)	3(6/7)	2(2/7)	2(3/7)	0(7/7)	0(7/7)	0(6/7)	0(7/7)
E17	1(2/5)	1(0/5)	1(0/5)	1(3/5)	1(4/5)	1(3/5)	4(1/5)	3(5/5)	2(2/5)	2(4/5)	0(5/5)	0(5/5)	0(5/5)	0(5/5)
E18	0(6/6)	0(6/6)	2(5/6)	1(0/6)	0(6/6)	1(4/6)	4(1/6)	2(5/6)	2(2/6)	1(4/6)	0(6/6)	0(6/6)	0(6/6)	0(6/6)
E19	1(1/8)	0(8/8)	2(3/8)	2(2/8)	1(1/8)	2(7/8)	1(0/8)	2(6/8)	1(4/8)	2(5/8)	0(8/8)	0(8/8)	0(7/8)	0(8/8)
E20	1(0/8)	0(8/8)	2(5/8)	0(5/8)	1(2/8)	1(5/8)	4(2/8)	1(8/8)	1(3/8)	2(4/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)

Note. [1]Abbreviations used in the table: R1 = Word Recognition Required; R2 = Vocabulary Knowledge Required; R3 = Syntactic Knowledge Required; R4 = Knowledge of Discourse Structure Required; R5 = Synthesis Required; R6 = Inference Required; R7 = Purpose of Information; T1 = Location of Information; T2 = Type of Match; T3 = Number of Plausible Distractors; T4 = Topical Knowledge Required; TW1 = Item Cues; TW2 = Guessing; TW3 = Surface Feature of Options.

Table 11

Consistencies between Verbal Data and Rating Data (Form F)

Item	R1[1]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3
F1	1(3/8)	0(8/8)	1(3/8)	1(2/8)	1(3/8)	0(8/8)	1(3/8)	2(5/8)	3(0/8)	0(3/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)
F2	2(1/7)	1(0/7)	2(4/7)	1(2/7)	1(0/7)	1(5/7)	1(0/7)	1(7/7)	1(3/7)	0(4/7)	0(7/7)	0(7/7)	0(7/7)	0(7/7)
F3	1(4/8)	0(8/8)	2(7/8)	0(8/8)	0(8/8)	0(5/8)	1(1/8)	2(8/8)	2(2/8)	1(2/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)
F4	1(4/6)	1(0/6)	1(0/6)	0(6/6)	0(6/6)	1(5/6)	4(0/6)	3(5/6)	3(2/6)	3(3/6)	0(6/6)	0(6/6)	0(5/6)	0(6/6)
F5	1(1/8)	0(8/8)	1(3/8)	0(8/8)	1(0/8)	0(8/8)	3(3/8)	3(8/8)	2(3/8)	0(5/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)
F6	1(2/7)	0(7/7)	1(0/7)	2(2/7)	2(4/7)	2(5/7)	1(2/7)	4(4/7)	3(2/7)	2(5/7)	0(7/7)	0(7/7)	0(5/7)	0(7/7)
F7	0(8/8)	0(8/8)	1(0/8)	1(2/8)	1(3/8)	2(6/8)	4(1/8)	4(6/8)	2(2/8)	1(6/8)	0(6/8)	0(8/8)	0(7/8)	0(8/8)
F8	2(0/7)	1(0/7)	2(0/7)	2(1/7)	1(3/7)	1(5/7)	5(0/7)	2(6/7)	2(3/7)	1(2/7)	0(7/7)	1(7/7)	0(7/7)	1(7/7)
F9	2(0/7)	0(7/7)	2(0/7)	2(1/7)	2(1/7)	1(0/7)	5(0/7)	1(4/7)	2(2/7)	2(4/7)	0(7/7)	0(7/7)	0(7/7)	0(7/7)
F10	2(2/7)	2(0/7)	2(1/7)	1(0/7)	1(0/7)	1(0/7)	4(0/7)	3(7/7)	2(2/7)	1(4/7)	0(6/7)	0(7/7)	0(7/7)	0(7/7)
F11	2(0/7)	1(0/7)	1(0/7)	2(4/7)	2(4/7)	2(0/7)	3(4/7)	4(7/7)	3(0/7)	1(6/7)	0(7/7)	0(7/7)	0(7/7)	0(7/7)
F12	1(3/6)	2(0/6)	2(3/6)	1(1/6)	1(0/6)	2(4/6)	4(1/6)	1(6/6)	2(3/6)	3(3/6)	0(5/6)	0(6/6)	0(6/6)	0(6/6)
F13	2(1/6)	1(0/6)	2(0/6)	1(2/6)	2(1/6)	2(5/6)	4(1/6)	3(3/6)	2(2/6)	3(4/6)	0(5/6)	0(6/6)	0(5/6)	0(6/6)
F14	2(4/5)	1(0/5)	2(5/5)	2(0/5)	2(0/5)	2(0/5)	4(0/5)	2(5/5)	1(4/5)	3(3/5)	0(3/5)	0(5/5)	0(4/5)	0(5/5)
F15	2(0/4)	2(0/4)	2(1/4)	1(0/4)	2(1/4)	2(0/4)	4(2/4)	3(3/4)	2(4/4)	3(3/4)	0(3/4)	0(4/4)	0(4/4)	0(4/4)
F16	1(0/8)	1(0/8)	1(0/8)	1(7/8)	2(8/8)	1(0/8)	3(5/8)	3(0/8)	3(0/8)	0(5/8)	0(8/8)	0(8/8)	0(8/8)	0(8/8)
F17	1(0/4)	0(4/4)	2(3/4)	0(4/4)	0(3/4)	0(4/4)	4(0/4)	3(3/4)	2(2/4)	2(1/4)	0(3/4)	0(4/4)	0(4/4)	0(4/4)
F18	2(2/6)	2(0/6)	2(4/6)	1(1/6)	1(0/6)	2(3/6)	2(0/6)	2(6/6)	2(1/6)	2(3/6)	0(5/6)	0(6/6)	0(4/6)	0(6/6)
F19	2(2/5)	1(0/5)	1(4/5)	2(0/5)	2(2/5)	2(0/5)	2(1/5)	1(3/5)	2(1/5)	2(3/5)	0(4/5)	0(5/5)	0(4/5)	0(5/5)
F20	2(5/7)	1(0/7)	2(4/7)	0(7/7)	1(0/7)	1(3/7)	1(0/7)	3(6/7)	1(1/7)	3(5/7)	0(5/7)	0(7/7)	0(6/7)	0(7/7)

Note. [1] Abbreviations used in the table: R1 = Word Recognition Required; R2 = Vocabulary Knowledge Required; R3 = Syntactic Knowledge Required; R4 = Knowledge of Discourse Structure Required; R5 = Synthesis Required; R6 = Inference Required; R7 = Purpose of Information; T1 = Location of Information; T2 = Type of Match; T3 = Number of Plausible Distractors; T4 = Topical Knowledge Required; TW1 = Item Cues; TW2 = Guessing; TW3 = Surface Feature of Options.

Using the criteria for distinguishing the levels of consistency between the rating data and the verbal report data on a given process for a given item, the frequency of occurrence within each level was examined. Tables 12 and 13 summarize for Form E and Form F, respectively, the number of items for which different levels of consistencies occurred between the two data sources. In both tables, the first row contains the 14 cognitive elements, and the first column contains different levels of consistency. As mentioned earlier, the two data sources were considered highly consistent when the ratio scores (i.e., the number of students reported the use of that item feature to arrive at the correct response in relation to the total number of students who correctly answered that item) > 0.5; moderately consistent when the ratio scores >0.25 and ≤ 0.5; low consistent when the ratio scores ≤ 0.25 and >0; and not consistent when the ratio scores = 0.

Table 12

Frequencies of Occurrence within Each Level of Consistency (Form E)

Process	R1[1]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3	Total
Highly Consistent	7	3	5	14	12	12	1	20	4	8	20	20	20	20	166
Moderately consistent	2	–	6	–	–	4	1	–	10	9	–	–	–	–	32
Low consistent	4	1	1	4	6	–	5	–	2	2	–	–	–	–	25
Not consistent	7	16	8	2	2	4	13	–	4	1	–	–	–	–	57

Note. [1] Abbreviations used in the table: R1 = Word Recognition Required; R2 = Vocabulary Knowledge Required; R3 = Syntactic Knowledge Required; R4 = Knowledge of Discourse Structure Required; R5 = Synthesis Required; R6 = Inference Required; R7 = Purpose of Information; T1 = Location of Information; T2 = Type of Match; T3 = Number of Plausible Distractors; T4 = Topical Knowledge Required; TW1 = Item Cues; TW2 = Guessing; TW3 = Surface Feature of Options.

[2] The symbol "–"means "Not applicable".

Table 13

Frequencies of Occurrence within Each Level of Consistency (Form F)

Process	R1[1]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3	Total
Highly Consistent	4	7	7	7	6	11	2	18	2	13	20	20	20	20	157
Moderately consistent	7	–	3	3	4	2	4	1	10	5	–	–	–	–	39
Low consistent	3	–	2	6	3	–	5		5	2	–	–	–	–	26
Not consistent	6	13	8	4	7	7	9	1	3	–	–	–	–	–	58

Note. [1] Abbreviations used in the table: R1 = Word Recognition Required; R2 = Vocabulary Knowledge Required; R3 = Syntactic Knowledge Required; R4 = Knowledge of Discourse Structure Required; R5 = Synthesis Required; R6 = Inference Required; R7 = Purpose of Information; T1 = Location of Information; T2 = Type of Match; T3 = Number of Plausible Distractors; T4 = Topical Knowledge Required; TW1 = Item Cues; TW2 = Guessing; TW3 = Surface Feature of Options.

[2] The symbol "–"means "Not applicable".

As shown in Tables 12 and 13, of a total of 560 (40 items x 14 variables) features coded for the reading items included in Forms E and F, 445 (79.5%) demonstrated different levels of consistency with the students' verbal reports of the cognitive processes they actually used to correctly answer the reading test items. This finding indicated that the item features were, overall, reasonably rated. Hence, no further modifications were made to the final set of consensus item ratings. Further examination of the two tables reveals that for both forms, high consistency between the item ratings and the verbal report data occurred most frequently for the cognitive elements T1 (Location of Information), T4 (Topical Knowledge Required), TW1 (Item Cues), TW2 (Guessing), and TW3 (Surface Feature of Options). The ratings for almost all items included in both forms for these five cognitive elements were supported by the verbal report data. The finding that the item ratings on T1 could be validated using the verbal report data was not surprising, considering that the students were instructed to verbalize the place where they were attending to while giving both concurrent and retrospective verbal reports. The findings that the ratings on T4, TW1, TW2, and TW3 showed high consistency with the verbal data were not surprising neither, considering that these cognitive processes were irrelevant to the constructs assessed by the MELAB reading and that this group of students with advanced-level English proficiency most probably do not need to arrive at answers using these "shortcuts".

Tables 12 and 13 also show that for both forms, the discrepancies between the item ratings and the students' verbal reports most frequently occurred on the elements R2 (Vocabulary Knowledge Required) and R7 (Purpose of Information). As can be seen from the last row of Table 12, for R2, 16 items (out of 20) did not show any consistency with the verbal report data, and for R7, 13 items did not show any consistency. Likewise, as seen from the last row of Table 13, for the features R2 and R7, 13 and 9 items did not show any consistency with the verbal report data, respectively. Considering that R2 and R7, as basic language knowledge and competence, are likely to become automatic for the students who are proficient ESL learners provides an explanation of why these processes were harder to trace through verbal description.

Implications for the Model Revision

The students' verbal reports provided valuable insights into the processes they used to correctly answer the MELAB reading test items and had several implications for revising the processing components in the initial cognitive model. First, the cognitive processes inferred from the students' verbal reports provided evidence that correctly answering an item often involved multiple processes, and that these processes covered 11 components of the proposed cognitive model: R1 (Word recognition), R3 (Using syntactic knowledge), R4 (Using knowledge of discourse structure), R5 (Synthesis), R6 (Drawing inferences), R7 (Using pragmatic knowledge), T1 (Locating specific details in text), T2 (Matching question stem to text), T3 (Evaluating alternative choices), T4 (Using topical knowledge), and TW2 (Guessing). Second, three components in the initial cognitive model — R2 (Using vocabulary knowledge), TW1 (Using clues in other items), and TW3 (Using surface features of answer choices) — were rarely or never reported by the students for the correct responses. Hence, these three components might be deleted from the initial model. Third, a new category, Using metacognitive strategies (T5), emerged from the verbal report data. T5 may need to be added to the initial model.

The verbal protocol results and the rater results regarding the model components were compared, and the outcomes of this comparison are presented in Table 14. An examination of the table reveals three points. First, the elements R1, R3, R4, R5, R6, R7, T1, T2, and T3 were found in both sets of results. Thus, these 9 elements were retained in the revised cognitive model. Second, the elements TW1 and TW3 were seldomly or never reported by the students for the correct responses, nor coded by the raters as relevant to reaching the correct responses. Hence, these two elements were deleted from the initial cognitive model. Third, the discrepancies between the two sets of results occurred for the elements R2, T4, TW2, and T5.

Table 14

A Comparison of the Verbal and Rater Results Regarding the Model Components

Item	R1[1]	R2	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3	T5
Rater	✓[2]	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	×
Verbal	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	×	✓

Note. [1] Abbreviations used in the table: R1 = Using word recognition; R2 = Using vocabulary knowledge; R3 = Using syntactic knowledge; R4 = Using knowledge of discourse structure; R5 = Synthesis; R6 = Drawing inferences; R7 = Recognizing authors' pragmatic or rhetorical purposes; T1 = Locating specific details in text; T2 = Matching question to text; T3 = Evaluating alternative options; T4 = Using topical knowledge; TW1 = Using item cues; TW2 = Guessing; TW3 = Using surface features of answer choices; T5 = Using metacognitive strategies.

[2] ✓ means the element can be obtained from the rater or verbal data; × means the element cannot be obtained from the rater or verbal data.

R2 was considered by the raters as needed to correctly answer the items, but was not found in the student protocols. As mentioned earlier, R2 likely was an automatic process that was used but seldom reported by this group of highly proficient students. In this sense, the raters' perspective of the cognitive processes required to correctly answer the items revealed automatic processes that could not be verbalized by the students. Further, prior research in L2 reading revealed a close relationship between R2 and reading comprehension (e.g., Bachman et al., 1996; Carr, 2003). Therefore, R2 was retained in the revised cognitive model.

T4 and TW2 were coded by the raters as not required to correctly answer the items, but reported mainly by two students as being used to arrive at the correct responses to a small number of items. Hence, the use of T4 and TW2 was considered as idiosyncratic ways of answering the reading test items and not representative of all the MELAB target examinees. Further, as mentioned earlier, Form E had more students who lacked topic familiarity than Form F, but the students' performances on the two forms did not show any significant difference. It appeared that topic familiarity was not that important for these two forms. As commented by one student (ES6), "…topic familiarity is not an important issue to pursue, and indeed, … topic familiarity is a very ambiguous term." The verbal data categorized as TW2 mainly reflected using vague hunches and intuition. For example, one student (FS2) verbalized, "I can't tell exactly why I chose B. I just feel that is correct answer." It appeared that TW2 represented cognitive processes that could not be exactly described by the students. Thus, T4 and TW2 were deleted from the initial cognitive model.

T5 was an additional component identified in the students' responses, but it was not identified by the raters. These proficient readers were good at using and describing the metacognitive strategies they used for planning, monitoring, and regulating their item solving processes. However, it was not clear whether or not T5 should be included in the final cognitive model as a new component and whether or not T5 affected the MELAB reading test item performance. Consequently, two revised cognitive models were obtained: One revised model contained 10 components – R1, R2, R3, R4, R5, R6, R7, T1, T2, and T3 – and the second model contained these 10 components plus T5. To examine for theoretical and statistical plausibility of developing an item difficulty model and to allow further evidence regarding the possible inclusion of T5, all these 11 components were submitted for the TBR analyses. The TBR analyses using 11 components as predictors for Form E and Form F are presented in the next chapter.

CHAPTER 6: ITEM DIFFICULTY MODELING USING TBR

The method and results for the third stage of the model validation are presented in two sections in this chapter. The first section describes the method for this phase. A description of the data is presented first. The procedure used for data analyses is then presented. The results are presented in the second section. The psychometric characteristics of the two test forms used for the current study are provided first, followed by the results from the TBR analyses of the two revised cognitive models obtained at the end of Phase 2 (i.e., the 10- and 11-component cognitive models) for Form E and Form F.

Method

Data

Two data files containing examinee item responses on Form E and Form F of the MELAB reading tests were provided by the ELI-UM. One data file contained the item responses from 1,703 examinees on Form E administered from January 2003 through September 2004. The second data file contained the item responses from 1,044 examinees on Form F administered from January 2003 through October 2004. The examinees who did not attempt one or more of the reading items (3.2% of the total number of examinees) were excluded by the ELI-UM under the assumption that such examinees were incapable of instigating the processes required to correctly answer the items (J. Jeffrey, personal communication, January 18, 2005). Consequently, neither file contained missing data.

Data Analysis Procedure

Data scoring and analysis of psychometric characteristics. Examinee response data were exported to the SPSS Version 13.0 (SPSS, Inc., 2005). Items were scored to the key provided by the ELI-UM, with 0 representing an incorrect response and 1 representing a correct response. The two dichotomously-scored datasets provided the basis for analyzing the psychometric characteristics of the test forms and calibrating the item parameter estimates. To better understand the two forms of the MELAB reading test used in this study, descriptive statistics and the reliabilities of Form E and Form F were computed.

Estimation of item parameters. Given the lack of local item independence due to common passages (Kolen & Brennan, 2004), item parameter estimates were calibrated using the testlet response theory (TRT) model (Wang, Bradlow, & Wainer, 2002). The TRT is a four-parameter dichotomous IRT model that introduces a testlet effect parameter, $\gamma_{ig(j)}$. The TRT model is expressed as:

$$p(y_{ij} = 1 \mid \theta_i) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j - \gamma_{ig(j)})]}{1 + \exp[a_j(\theta_i - b_j - \gamma_{ig(j)})]}, \qquad (6)$$

where y_{ij} is the score for examinee i on item j, θ_i is the ability level of examinee i, $p(y_{ij} = 1|\theta_i)$ is the probability that examinee i with ability θ_i will correctly answer item j, a_j is the discrimination parameter of item j, b_j is the difficulty parameter of item j, c_j is the pseudo-guessing parameter of item j, and $\gamma_{ig(j)}$ is the testlet effect parameter indicating the testlet effect for examinee i responding to item j that is nested in testlet g. As Equation 6 shows, the TRT model separates the testlet effect from examinee ability by estimating the testlet effect parameter (γ) for each testlet and each examinee during the calibration of the *a*-, *b*-, and *c*-parameters. In this way, the problem of local dependence of the items referenced to the same reading passage is accounted for, and the resulting item parameter estimates are more accurate (Wang et al., 2002).

For the current study, the parameters of the reading items were estimated separately within Form E and Form F. The computer program SCORIGHT 3.0 (Wang, Bradlow, & Wainer, 2004), which is based on the TRT model, was used. The item difficulty parameter estimates obtained for each form were formatted into two 20 x 1 vectors, which were used as criterion for the TBR analyses.

TBR analyses. TBR analyses were performed to determine the extent to which the identified cognitive processes associated with each item explained the item difficulty. As mentioned at the end of the previous chapter, two item difficulty models were analyzed for each form. In the first analysis, the 10 components (i.e., R1, R2, R3, R4, R5, R6, R7, T1, T2, and T3) that were corroborated by the raters and inferred from the students' verbal protocols were used as predictors of item difficulty. In the second analysis, the 10 predictors used in the first analysis and T5, which was identified in the students' verbal protocols but not by the raters, were used as predictors of item difficulty.

Given the nature of the scale data (see Table 1) and verbal report data, the scale data were used as predictors in the TBR analyses. Specifically, as not all cognitive processes were reported or reported every time by the students when they were used, it was not clear whether the non-verbalized processes were indeed not used by the students. For example, for the students in the current study, the use of some of the processes like R2 was likely to be automatic rather than controlled. Further, the verbal report data were not sensitive enough to discern the levels for some of the variables. For instance, for the variable R3 (Syntax Knowledge Required), it was not possible to discern in the students' protocols whether it was somewhat or critically used by the students to correctly answer an item. In contrast, the raters were able to distinguish the cognitive processes involved in item solving more finely. Therefore, the scales provided in Table 1 and corroborated by the raters were retained for the TBR analyses. For the variable T5 that was not identified by the raters, frequency ratios were used, that is, the number of students reporting the use of this process for the correct responses to a given item in relation to the total number of students who correctly answered that item.

The Classification and Regression Tree (CRT) module provided as part of SPSS 13.0 (SPSS, Inc., 2005) was used to conduct the TBR analyses. The aim of the CRT is to maximize within-node homogeneity with respect to the dependent variable. Using the CRT, a maximal tree is grown that classifies all cases into its own cluster or terminates when one or more predetermined limits are imposed by the user, for example, when the number of cases in all terminal nodes falls below five cases per node (Rupp et al., 2001). In the tree evolution, predictor variables can be used repeatedly for splits, and the best split is the one that results in the largest reduction between the impurity of the parent node and the sum of the impurity of the two child nodes. Impurity is a statistical measure of the dissimilarity in the dependent variable among the items within a single node. The smaller the impurity value is, the more homogeneous the items within a node. The CRT uses improvement values to measure the decrease in impurity caused by the split of a given node (AnswerTree User's Guide, 2002).

The TBR analysis of Form E with 10 predictors began with the placement of the 20 items in a single cluster (i.e., node) at the top of the tree, where 0% of the variance was explained. Using the 10 components as predictors, the items were successively split into nodes that were as homogeneous as possible with respect to item difficulty. A binary recursive partitioning algorithm (Breiman et al., 1984) was used to evaluate all possible splits of the predictor variables at each level of splitting. Using this binary algorithm, clusters were split into two nodes at each iteration and the nodes in turn become parent nodes and were split into the left and right child nodes, wherever possible. The best split was determined using the improvement values. After finding the optimal split, the CRT was repeated until one or more of the stopping rules were met. Since a small set of items was used, the stopping rules were: (1) an improvement value was less than 0.0001; (2) the parent node had fewer than three items; and (3) the parent node, if split, would result in a child node with no items.

After the tree was grown, the optimal number of the tree levels and terminal nodes was determined based on the amount of variance in difficulty explained (R^2) and the parsimony and interpretability of the tree model (Ewing & Huff, 2004; Rupp et al., 2001). If adding a new level of splits and more terminal nodes did not lead to a relatively large improvement in the variance explained, the more parsimonious model was selected (Huff, 2003). For example, if a tree with 8 terminal nodes explained 79% of the variance and the tree with 9 terminal nodes explained 80% of the variance, the former would be selected as the optimal solution due to its parsimony. To achieve model parsimony and to improve model interpretability, pruning was conducted, which involved removing one or more sets of child nodes and creating a tree with the fewest possible terminal nodes. Pruning is subjective in nature. However, the final tree has to have an acceptable error rate. In this study, pruning was conducted when adding another level of split increased less than 2% of the variance explained (Huff, 2003). In sum, the final tree model should represent a reasonable compromise among variance explained, error, parsimony, and practical interpretability (i.e., making theoretical sense). As Rupp et al. (2001) commented, "any R^2 measure in a regression tree setting is dependent on the somewhat subjective selection of the tree that is interpreted" (p. 203).

Using the procedure specified above, the TBR analyses were first performed using the Form E data to test the cognitive model with 10 components, and then to test the cognitive model with 11 components. Then, the TBR analyses were conducted using the Form F data to determine whether the item difficulty models for Form F matched the models obtained for Form E. Consistencies between the two forms would provide strong empirical support for the final cognitive model of the MELAB reading test items.

Results

Psychometric Characteristics

The psychometric characteristics of Form E and Form F are summarized in Table 15.

The mean score of the 1,703 examinees for Form E was 10.94 (maximum score = 20), and the standard deviation was 4.19. The mean score of the 1,044 examinees for Form F was 10.71, and the standard deviation was 3.65. The mean scores on Form E and Form F were compared using the independent samples *t*-test. The results indicated that there was no significant difference between the two forms at the 0.05 level of significance ($t = 1.54$, $p > 0.05$). While the results of the *F* test for two independent samples indicated that the two forms did not have equal variances at the 0.05 level of significance, $F_{(1702, 1043)} = 1.32$, $p < 0.05$, the value of the *F*-ratio was close to one and the significance was attributed to the large degrees of freedom in the numerator. The range of the scores for Form E (20) was comparable to that for Form F (19). The medians of the two forms were identical ($Median_E = Median_F = 11$). The skewness for Form E was 0.10 and for Form F was 0.21, which indicated that the distributions of the scores for both forms were approximately symmetrical around the mean. The kurtosis for Form E was –0.70 and for Form F was –0.50, which indicated that the distributions of the scores for both forms were platykurtic. The reliability of Form F (0.71) was slightly lower than that of Form E (0.79). Notwithstanding the ambiguity of the homogeneity of variance test, these results indicated that the two forms were nominally parallel.

Table 15

Descriptive Statistics and Reliability for the Two Forms

Form	*Min.*	*Max.*	*Median*	*Mean*	*SD*	*Skew.*	*Kurto.*	*Reliability*
E	0.00	20.00	11.00	10.94	4.19	0.10	-0.70	0.79
F	1.00	20.00	11.00	10.71	3.65	0.21	-0.50	0.71

Figures 7 and 8 display the distributions of the total raw scores on Form E and Form F, respectively. As seen in the figures, both distributions are uni-modal, appear approximately symmetrical around the mean, and are somewhat flatter than the normal curve fit to the data.

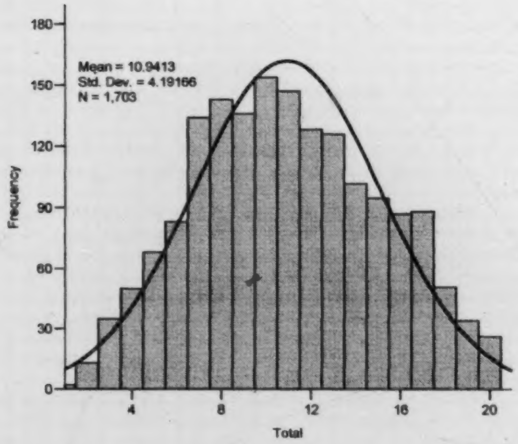


Figure 7. Histogram of total scores on Form E.

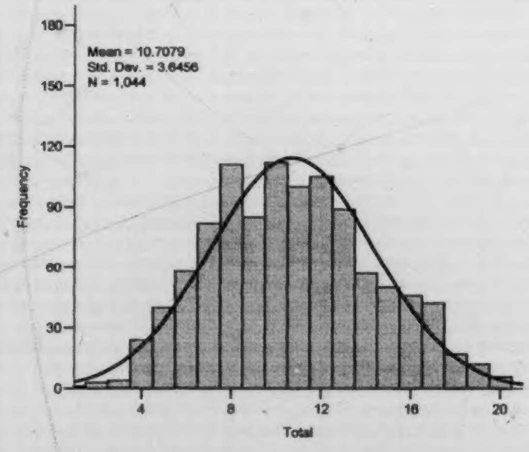


Figure 8. Histogram of total scores on Form F.

Figure 9 displays the test characteristic curves (TCCs) for Form E and Form F. The X-axis represents examinees' ability scale. The Y-axis represents examinees' expected total score for a given ability. As seen in the figure, for examinees with below-average ability, the differences in the expected total scores for a given ability between the two forms are slight. To determine whether the TCCs of the two forms were significantly different, the mean square residual (MSR) (Puhan, 2003) was calculated. The MSR result indicated that the difference between the TCCs of the two forms was not statistically significant (*MSR* = 1.025, *df* = 29, $p > 0.05$). Therefore, the two forms were regarded as parallel[1].

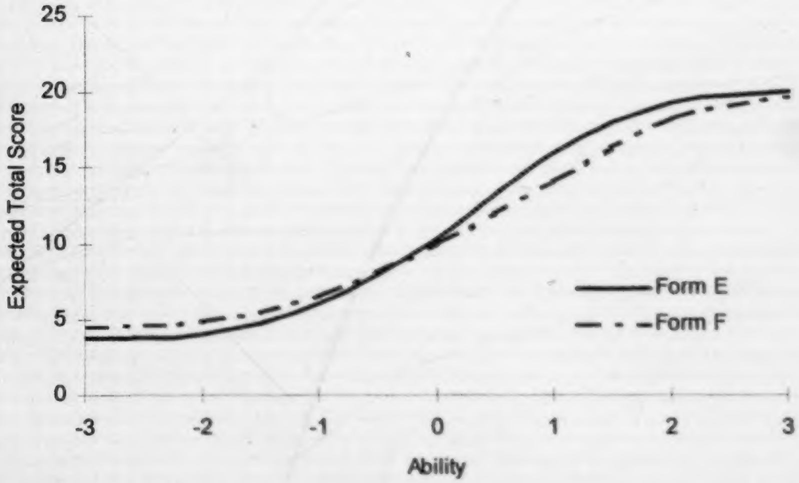


Figure 9. Test characteristic curves for Form E and Form F.

The item difficulty parameter estimates obtained for each form are presented in ascending order in Table 16. As shown in this table, the values of *b* ranged from –1.16 to 1.88 for Form E, with the mean 0.23 and the standard deviation 0.85. The values of *b* ranged from –1.16 to 1.78 for Form F, with the mean 0.42 and the standard deviation 0.99. While Form E had four difficult items (i.e., *b* values larger than 1) and Form F contained eight difficult items, the easy items (i.e., *b* values less than 0) for Form F were somewhat easier than the easy items for Form E. Consequently, as revealed earlier, the two forms were overall equally difficult.

Table 16

Item Difficulty Parameter Estimates for All Items

Item	Item Difficulty (*b*)	Item	Item Difficulty (*b*)
E9	-1.16	F5	-1.16
E13	-1.01	F16	-1.03
E4	-0.72	F8	-0.86
E3	-0.62	F1	-0.71
E7	-0.46	F2	-0.62
E2	-0.30	F10	-0.45
E6	-0.25	F3	-0.42
E11	-0.23	F9	0.09
E12	0.09	F11	0.43
E18	0.19	F6	0.47
E20	0.23	F19	0.54
E1	0.26	F7	0.84
E14	0.47	F4	1.11
E16	0.65	F14	1.11
E19	0.67	F17	1.37
E8	0.85	F12	1.39
E5	1.24	F13	1.48
E15	1.31	F20	1.53
E10	1.54	F18	1.57
E17	1.88	F15	1.78
Mean	0.23	*Mean*	0.42
SD	0.85	*SD*	0.99

TBR Results

As mentioned earlier, the two revised cognitive models obtained at the end of Phase 3 of the study were evaluated. The first model did not include T5, while the second did.

Hence, in the first set of TBR analysis, 10 predictors of item difficulty were used. In the second set of TBR analysis, 11 predictors of item difficulty were used, including T5. In this section, the results from two sets of the TBR analyses are first presented for Form E and then for Form F, followed by a comparison of the results for the two test forms.

Test of the 10-component cognitive model (Form E). All 10 predictors were included in the final tree for Form E: T3, R3, R1, R5, R4, T1, R2, T2, R7, and R6. Taken together, these ten variables accounted for 97.9% of the variance in item difficulty, which indicated that this was a good model. The tree diagram displayed in Figure 10 provides a graphic representation of the final tree solution for the 10-component cognitive model using the Form E data. The values for the variance explained at each level of split can be found to the left of the tree. Figure 10 also shows, in each node of the tree, the item means, and the number of items at that node. For example, Node 1 shows that the overall mean item difficulty was 0.232 and this node contained 20 items.

As Figure 10 shows, the TBR solution began with all 20 items in a single node with 0% variance explained. The first variable entering the tree was T3 (Number of Plausible Distractors), which separated the items into two nodes. The left node contained 8 items coded as having zero or one plausible distractor (*Item Mean* = -0.332). The right node contained 12 items coded as having two or three plausible distractors (*Item Mean* = 0.608). This split indicated that items with more than one plausible distractor were more difficult than items with zero or one plausible distractor.

Subsequently, the items within Node 2 were split into two nodes based on R3

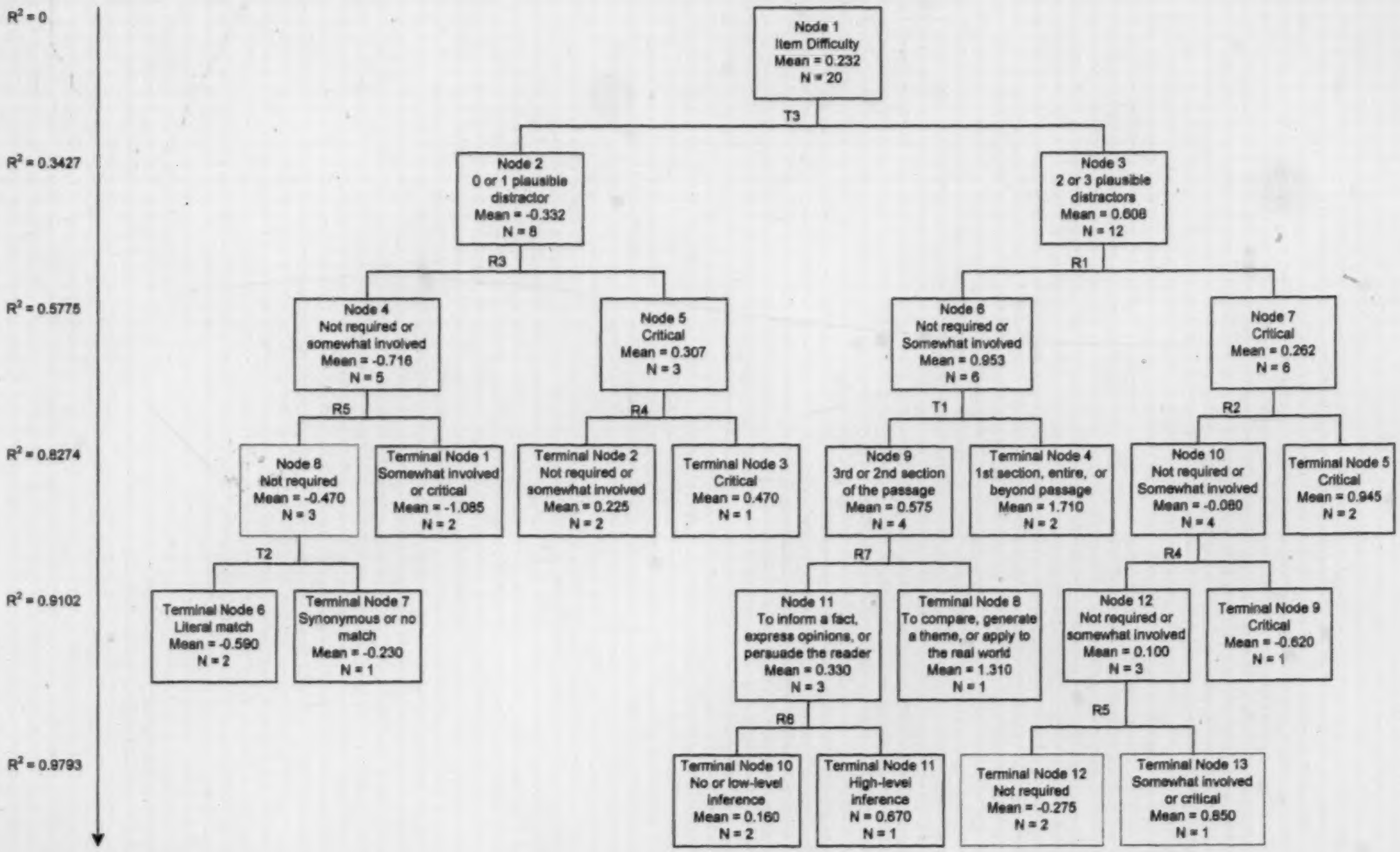

R² = 0
R² = 0.3427
R² = 0.5775
R² = 0.8274
R² = 0.9102
R² = 0.9793
Node 1 Item Difficulty Mean = 0.232 N = 20
T3
Node 2 0 or 1 plausible distractor Mean = -0.332 N = 8
Node 3 2 or 3 plausible distractors Mean = 0.608 N = 12
R3
R1
Node 4 Not required or somewhat involved Mean = -0.716 N = 5
Node 5 Critical Mean = 0.307 N = 3
Node 6 Not required or Somewhat involved Mean = 0.953 N = 6
Node 7 Critical Mean = 0.262 N = 6
R5
R4
T1
R2
Node 8 Not required Mean = -0.470 N = 3
Terminal Node 1 Somewhat involved or critical Mean = -1.085 N = 2
Terminal Node 2 Not required or somewhat involved Mean = 0.225 N = 2
Terminal Node 3 Critical Mean = 0.470 N = 1
Node 9 3rd or 2nd section of the passage Mean = 0.575 N = 4
Terminal Node 4 1st section, entire, or beyond passage Mean = 1.710 N = 2
Node 10 Not required or Somewhat involved Mean = -0.080 N = 4
Terminal Node 5 Critical Mean = 0.945 N = 2
T2
R7
R4
Terminal Node 6 Literal match Mean = -0.590 N = 2
Terminal Node 7 Synonymous or no match Mean = -0.230 N = 1
Node 11 To inform a fact, express opinions, or persuade the reader Mean = 0.330 N = 3
Terminal Node 8 To compare, generate a theme, or apply to the real world Mean = 1.310 N = 1
Node 12 Not required or somewhat involved Mean = 0.100 N = 3
Terminal Node 9 Critical Mean = -0.620 N = 1
R6
R5
Terminal Node 10 No or low-level inference Mean = 0.160 N = 2
Terminal Node 11 High-level inference N = 0.670 N = 1
Terminal Node 12 Not required Mean = -0.275 N = 2
Terminal Node 13 Somewhat involved or critical Mean = 0.850 N = 1

Figure 10. Tree diagram for the cognitive model with 10 components (Form E).

(Syntactic Knowledge Required). The five items coded as not or somewhat requiring syntactic knowledge were assigned to the left node (*Item Mean* = -0.716), and the three items coded as critically requiring syntactic knowledge were assigned to the right node (*Item Mean* = 0.307). This split indicated that items that required knowledge of complex infrequently used syntax were more difficult than items that did not. Next, the five items within Node 4 were further split based on R5 (Synthesis Required). The three items coded as not requiring synthesis were assigned to the left node (*Item Mean* = -0.470), and the two items coded as requiring low- or high-level synthesis were assigned to the right node (*Item Mean* = -1.085). Contrary to expectation, this split indicated that the items that required synthesis were easier than the items that did not. As the right node did not have any child nodes, it became a terminal node. The items within Node 8 were split based on T2 (Type of Match). The two items coded as requiring a literal match between question and text were assigned to the left node (*Item Mean* = -0. 590), and the one item coded as synonymous or no match between question and text was assigned to the right node (*Item Mean* = -0.230). The results at this split indicated that items with a question stem that synonymously matched or did not match the text were more difficult than the item with a question stem that literally matched the text. Due to the lack of child nodes, these two nodes were terminal nodes. Returning to Node 5, the three items in Node 5 were split based on R4 (Knowledge of Text Structure Required). The two items coded as not or somewhat requiring the knowledge of text structure were assigned to the left node (*Item Mean* = 0.225), and the single item coded as critically requiring the knowledge of text structure was assigned to the right node (*Item Mean* = 0.470). This split indicated that the item critically requiring knowledge of text structure were more difficult than the items not or somewhat requiring the knowledge of text structure. Due to the lack of child nodes, these two nodes were terminal nodes.

On the right side of the tree, the items coded as having two or three plausible distractors were first evenly split into two nodes based on R1 (Word Recognition Required). The left node of this split contained 6 items coded as not or somewhat requiring word recognition (*Item Mean* = 0.953), and the right node contained 6 items coded as critically requiring word recognition (*Item Mean* = 0.262). Contrary to expectation, this split indicated that the items that not or somewhat required word recognition were more difficult than the items that critically required word recognition. Subsequently, the items within Node 6 were further split into two nodes based on T1 (Location of Information). The four items requiring information located in the second or third section of the passage were assigned to the left node (*Item Mean* = 0.575), and the two items requiring information located in the first section of the passage, entire passage, or beyond the passage were assigned to the right node (*Item Mean* = 1.710). This split indicated that items requiring information in the earlier part of the passage, the entire passage, or beyond the passage were more difficult than items requiring information in the middle or later part of the passage. As the right node does not have child nodes, it became a terminal node.

Subsequently, the items within Node 9 were further split based on R7 (Purpose of

Information). The left node contained three items coded as to inform a fact, state a procedure or action, express opinions, or persuade the reader (*Item Mean* = 0.330), and the right node contained one item coded as to compare or contrast, generate a theme, or apply to the real world (*Item Mean* = 1.310). This split indicated that items that required examinees to compare or contrast, generate a theme, or apply to the real world were more difficult than items that required examinees to understand a fact, a procedure, an action, or the authors' opinions. As the right node did not have any child nodes, it became a terminal node. Lastly, the items within Node 11 were split into two terminal nodes based on R6 (Synthesis Required). The two items coded as requiring no inference or low-level inference were assigned to the left node (*Item Mean* = 0.160), and the one item coded as requiring high-level inference was assigned to the right node (*Item Mean* = 0.670). This split indicated that the item requiring high-level inference was more difficult than the items requiring no or low-level inference.

Returning to Node 7, the six items in Node 7 were split into two nodes based on R2 (Vocabulary Knowledge Required). The left node contained four items coded as not or somewhat requiring vocabulary knowledge (*Item Mean* = -0.080), and the right node contained two items coded as critically requiring vocabulary knowledge (*Item Mean* = 0.945). This split indicated that the items that required the knowledge of infrequently used and/or specialized vocabulary were more difficult than the items that did not. The right node was a terminal node in that no further split occurred. Next, the items within Node 10 were further split based R4 (Knowledge of Text Structure Required). The three items contained in the left node were coded as not or somewhat requiring knowledge of text structure (*Item Mean* = 0.100). The one item in the right node was coded as critically requiring knowledge of text structure (*Item Mean* = -0.620). Contrary to expectation, this split indicated that the items not or somewhat requiring knowledge of text structure were more difficult than the item critically requiring knowledge of text structure. Due to the lack of child nodes, the right node became a terminal node. Finally, the items within Node 12 were split based on R5 (Synthesis Required). The two items not requiring synthesis were assigned to the left node (*Item Mean* = -0.275), and the one item somewhat or critically requiring synthesis was assigned to the right node (*Item Mean* = 0.850). This split indicated that the item that required synthesis was more difficult than the items that did not.

The final tree for the 10-component cognitive model using the Form E data included all 10 components, produced 13 terminal nodes, and involved five levels of splits. The first split at the root node explained 34.3% of the variance in item difficulty. The two splits at the second level increased the amount of variance explained to 57.8%. The four splits at the third level increased the variance explained to 82.7%. The three splits at the third level increased the portion of variance explained to 91.0%, and the two splits at the last level increased the variance explained to 97.9%. Given the magnitude of the increase of the variance explained at each level of split, no pruning was necessary.

Test of the 11-component cognitive model (Form E). The TBR analysis of the cognitive model with 11 components using the Form E data produced an initial tree including nine predictors: eight of the 10 predictors in the Form E tree for the 10-compoent cognitive model (i.e., T3, R3, R1, R4, T1, R2, R7, and R6), plus T5 (Metacognitive Strategies Used). This tree produced 12 terminal nodes and accounted for 93.0% of the variance in item difficulty (see Appendix K). However, when the two terminal nodes produced by the split based on R6 were pruned, the resulting tree had 11 terminal nodes and explained 91.9% of the variance. Given the small change in variance explained and the desire for parsimony, the latter tree containing eight predictors (T3, R3, R1, R4, T1, R2, R7, and T5) was selected as the final tree for the 11-component cognitive model using the Form E data (see Figure 11).

As seen in Figure 11, the first variable entering the tree was T3 (Number of Plausible Distractors), which separated the items into two nodes. The left node contained 8 items coded as having zero or one plausible distractor (*Item Mean* = -0.332), while the right node contained 12 items coded as having two or three plausible distractors (*Item Mean* = 0.608). This split indicated that items with more than one plausible distractor were more difficult than items with zero or one plausible distractor.

Subsequently, the items within Node 2 were split into two nodes based on R3 (Syntactic Knowledge Required). The five items coded as not or somewhat requiring syntactic knowledge were assigned to the left node (*Item Mean* = -0.716), and the three items coded as critically requiring syntactic knowledge were assigned to the right node (*Item Mean* = 0.307). This split indicated that items that required knowledge of complex infrequently used syntax were more difficult than items that did not. Next, the five items

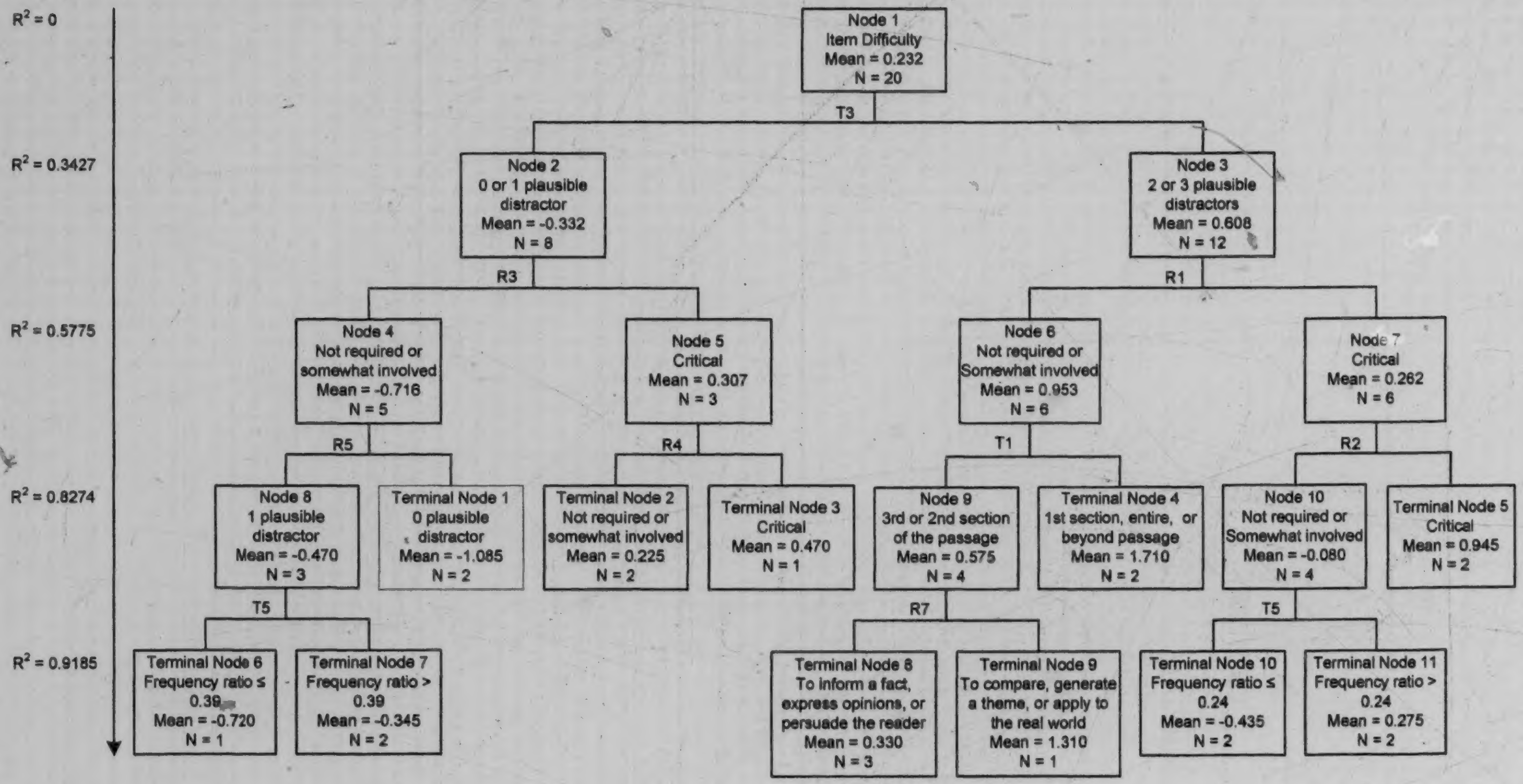

R² = 0
R² = 0.3427
R² = 0.5775
R² = 0.8274
R² = 0.9185
Node 1
Item Difficulty
Mean = 0.232
N = 20
T3
Node 2
0 or 1 plausible distractor
Mean = -0.332
N = 8
Node 3
2 or 3 plausible distractors
Mean = 0.608
N = 12
R3
R1
Node 4
Not required or somewhat involved
Mean = -0.716
N = 5
Node 5
Critical
Mean = 0.307
N = 3
Node 6
Not required or Somewhat involved
Mean = 0.953
N = 6
Node 7
Critical
Mean = 0.262
N = 6
R5
R4
T1
R2
Node 8
1 plausible distractor
Mean = -0.470
N = 3
Terminal Node 1
0 plausible distractor
Mean = -1.085
N = 2
Terminal Node 2
Not required or somewhat involved
Mean = 0.225
N = 2
Terminal Node 3
Critical
Mean = 0.470
N = 1
Node 9
3rd or 2nd section of the passage
Mean = 0.575
N = 4
Terminal Node 4
1st section, entire, or beyond passage
Mean = 1.710
N = 2
Node 10
Not required or Somewhat involved
Mean = -0.080
N = 4
Terminal Node 5
Critical
Mean = 0.945
N = 2
T5
R7
T5
Terminal Node 6
Frequency ratio ≤ 0.39
Mean = -0.720
N = 1
Terminal Node 7
Frequency ratio > 0.39
Mean = -0.345
N = 2
Terminal Node 8
To inform a fact, express opinions, or persuade the reader
Mean = 0.330
N = 3
Terminal Node 9
To compare, generate a theme, or apply to the real world
Mean = 1.310
N = 1
Terminal Node 10
Frequency ratio ≤ 0.24
Mean = -0.435
N = 2
Terminal Node 11
Frequency ratio > 0.24
Mean = 0.275
N = 2

Figure 11. Tree diagram for the cognitive model with 11 components (Form E).

within Node 4 were further split based on R5 (Synthesis Required). The three items coded as not requiring synthesis were assigned to the left node (*Item Mean* = -0.470), and the two items coded as requiring low- or high-level synthesis were assigned to the right node (*Item Mean* = -1.085). Again, contrary to expectation, this split indicated that the items that required synthesis were easier than the items that did not. As the right node did not have any child nodes, it became a terminal node. The items within Node 8 were then split into two terminal nodes based on T5 (Metacognitive Strategies Used). The items for which metacognitive strategies were less frequently used (frequency ratios ≤ 0.39) were assigned to the left node (*Item Mean* = -0.720), and the items for which metacognitive strategies were more frequently used (frequency ratios > 0.39) were assigned to the right node (*Item Mean* = -0.345). This indicated that items that required the use of metacognitive strategies were more difficult than items that did not. Returning to Node 5, the three items in Node 5 were split based on R4 (Knowledge of Text Structure Required). The two items coded as not or somewhat requiring knowledge of text structure were assigned to the left node (*Item Mean* = 0.225), and the single item coded as critically requiring knowledge of text structure was assigned to the right node (*Item Mean* = 0.470). This split indicated that the item that critically required the knowledge of text structure was more difficult than the items that did not or somewhat required the knowledge of text structure. Due to the lack of child nodes, these two nodes became terminal nodes.

On the right side of the tree, the items coded as having two or three plausible distractors were first evenly split into two nodes based on R1 (Word Recognition

Required). The left node of this split contained 6 items coded as not or somewhat requiring the process of R1 (*Item Mean* = 0.953), and the right node contained 6 items coded as critically requiring the process of R1 (*Item Mean* = 0.262). Counter-intuitively, this split indicated that the items that critically required word recognition were easier than the items that did not or somewhat required word recognition. Subsequently, the items within Node 6 were further split into two nodes based on T1 (Location of Information). The four items requiring information located in the second or third section of the passage were assigned to the left node (*Item Mean* = 0.575), and the two items requiring information located in the first section of the passage, entire passage, or beyond the passage were assigned to the right node (*Item Mean* = 1.710). This split indicated that items requiring information in the earlier part of the passage, entire passage, or beyond the passage were more difficult than items requiring information in the middle or later part of the passage. As the right node did not have child nodes, it became a terminal node.

Subsequently, the items within Node 9 were further split into two terminal nodes based on R7 (Purpose of Information). The left node contained three items coded as to inform a fact, state a procedure or action, express opinions, or persuade the reader (*Item Mean* = 0.330). The right node contained one item coded as to compare or contrast, generate a theme, or apply to the real world (*Item Mean* = 1.310). This split indicated that items that required examinees to compare or contrast, generate a theme, or apply to the real world were more difficult than items that required examinees to understand a fact, a procedure, an action, or the authors' opinions.

Returning to Node 7, the six items in Node 7 were split into two nodes based on R2 (Vocabulary Knowledge Required). The left node contained four items coded as not or somewhat requiring vocabulary knowledge (*Item Mean* = -0.080), and the right node contained two items coded as critically requiring vocabulary knowledge (*Item Mean* = 0.945). This split indicated that items that required the knowledge of infrequently used and/or specialized vocabulary were more difficult than items that did not. The right node was a terminal node in that no further split occurred. Finally, the items within Node 10 were split into two terminal nodes based on T5 (Metacognitive Strategies Used). The two items for which metacognitive strategies were less frequently used (frequency ratios ≤ 0.24) were assigned to the left terminal node (*Item Mean* = -0.435), and the two items for which such strategies were more frequently used (frequency ratios > 0.24) were assigned to the right terminal node (*Item Mean* = 0.275). This split indicated that items that required the use of metacognitive strategies were more difficult than items that did not.

The final tree for the 11-component cognitive model using the Form E data included eight components, produced 11 terminal nodes, and involved four levels of splits. The first split, at the root node, explained 34.3% of the variance in item difficulty. The two splits at the second level increased the amount of variance explained to 57.8%. The four splits at the third level increased the variance explained to 82.7%. The three splits at the last level increased the portion of variance explained to 91.9%.

A comparison of the 10- and 11-component cognitive model (Form E). A comparison of the final tree solutions for the 10- and 11-component cognitive model based on the

Form E data, as shown in Figures 10 and 11, revealed four major points. First, for the 10-component cognitive model, the final tree included all 10 components in the cognitive model. However, for the 11-component model, the final tree included nine components in the cognitive model; R6 and T2 were not included in the final tree. Second, eight predictors – T3, R3, R1, R5, R4, T1, R2, and R7 – were in the same positions and produced the same results of split for both the 10- and 11-component cognitive models. However, the items within Node 8 and Node 10 were split, respectively, by T2 (Type of Match) and R4 (Knowledge of Discourse Structure Required) in the 10-component model, and both by T5 (Metacognitive Strategies Used) in the 11-component model. Third, the final tree for the 10-component cognitive model had five levels of split. However, following pruning, the final tree for the 11-component cognitive model had four levels of split. Lastly, the 10-component item difficulty model explained 97.9% of the variance in item difficulty, but the 11-component item difficulty model, after pruning, explained 91.9% of the variance in item difficulty. Hence, the inclusion of T5 in the cognitive model did not substantially increase the explanatory power of the item difficulty model. Given the amount of variance explained and the statistical principle of parsimony (Kerlinger, 1979), it appeared that the cognitive model without T5 is a better model. Overall, the TBR results for Form E supported the 10-component cognitive model.

Test of the 10-component cognitive model (Form F). The TBR analysis of the cognitive model with 10 components using the Form F data produced an initial tree that included nine predictors: T3, R6, R4, R3, R1, T2, T1, R7, and R2. This tree produced 13 terminal nodes and accounted for 99.4% of the variance in item difficulty (see Appendix L). However, when the two terminal nodes produced by the split based on T1 were pruned, the resulting tree had 12 terminal nodes and explained 99.3% of the variance. Given the small change in variance explained and the desire for parsimony, the latter tree containing nine predictors (T3, R6, R4, R3, R1, T2, T1, R7, and R2) was selected as the final tree for the 10-component cognitive model using the Form F data (see Figure 12).

As seen in Figure 12, the TBR solution began with all 20 items in a single node with 0% variance explained. The first variable entering the tree was, again, T3 (Number of Plausible Distractors), which separated the 20 items into two nodes. The left node contained 9 items coded as having zero or one plausible distractor (*Item Mean* = -0.442). The right node contained 11 items coded as having two or three plausible distractors (*Item Mean* = 1.131). This split again indicated that items with more than one plausible distractor were more difficult than items with zero or one plausible distractor.

The items within Node 2 were then split into two nodes based on R6 (Inference Required). The left node contained seven items coded as requiring no or low-level inference (*Item Mean* = -0.750), and the right node contained two items coded as requiring high-level inference (*Item Mean* = 0.635). This split indicated that the item that required high-level inferences were more difficult than the items that required no or low-level inferences. As the right node did not have any child nodes, it became a terminal node. Next, the seven items within Node 4 were further split into two nodes based on R3 (Syntactic Knowledge Required). The left node contained three items coded as not or

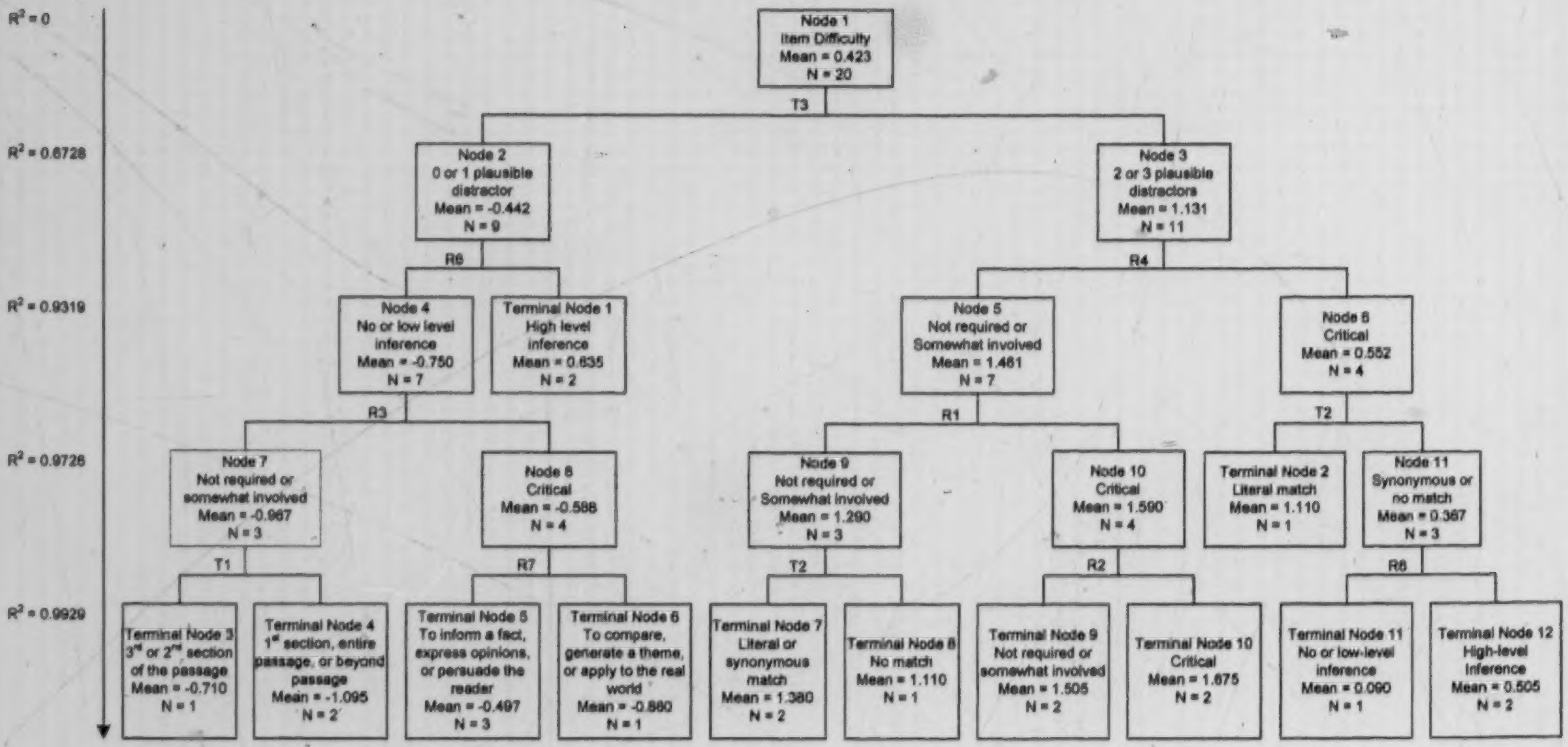

R² = 0
R² = 0.6728
R² = 0.9319
R² = 0.9726
R² = 0.9929
Node 1 Item Difficulty Mean = 0.423 N = 20
T3
Node 2 0 or 1 plausible distractor Mean = -0.442 N = 9
Node 3 2 or 3 plausible distractors Mean = 1.131 N = 11
R6
R4
Node 4 No or low level inference Mean = -0.750 N = 7
Terminal Node 1 High level inference Mean = 0.635 N = 2
Node 5 Not required or Somewhat involved Mean = 1.461 N = 7
Node 6 Critical Mean = 0.552 N = 4
R3
R1
T2
Node 7 Not required or somewhat involved Mean = -0.967 N = 3
Node 8 Critical Mean = -0.588 N = 4
Node 9 Not required or Somewhat involved Mean = 1.290 N = 3
Node 10 Critical Mean = 1.590 N = 4
Terminal Node 2 Literal match Mean = 1.110 N = 1
Node 11 Synonymous or no match Mean = 0.367 N = 3
T1
R7
T2
R2
R8
Terminal Node 3 3rd or 2nd section of the passage Mean = -0.710 N = 1
Terminal Node 4 1st section, entire passage, or beyond passage Mean = -1.095 N = 2
Terminal Node 5 To inform a fact, express opinions, or persuade the reader Mean = -0.497 N = 3
Terminal Node 6 To compare, generate a theme, or apply to the real world Mean = -0.860 N = 1
Terminal Node 7 Literal or synonymous match Mean = 1.380 N = 2
Terminal Node 8 No match Mean = 1.110 N = 1
Terminal Node 9 Not required or somewhat involved Mean = 1.505 N = 2
Terminal Node 10 Critical Mean = 1.675 N = 2
Terminal Node 11 No or low-level inference Mean = 0.090 N = 1
Terminal Node 12 High-level Inference Mean = 0.505 N = 2

Figure 12. Tree diagram for the cognitive model with 10 components (Form F).

somewhat requiring syntactic knowledge (*Item Mean* = -0.967), and the right node contained four items coded as critically requiring syntactic knowledge (*Item Mean* = -0.588). This split indicated that items that required the knowledge of complex infrequently used syntax were more difficult than items that did not. The three items within Node 7 were further split into two terminal nodes based on T1 (Location of Information). The single item coded as requiring information located in the second or third section of the passage was assigned to the left node (*Item Mean* = -0.710), and the two items coded as requiring information located in the first section of the passage, the entire passage, or beyond the passage were assigned to the right node (*Item Mean* = -1.095). Contrary to expectation, this split indicated that the item requiring information located in the second or third section of the passage was more difficult than the items requiring information located in the first section of the passage, the entire passage, or beyond the passage. Lastly, the items within Node 8 were further split into two terminal nodes based on R7 (Purpose of Information). The three items coded as to inform a fact, state a procedure or action, express opinions, or persuade the reader were assigned to the left node (*Item Mean* = -0.497), and the single item coded as to compare or contrast, generate a theme, or apply to the real world was assigned to the right node (*Item Mean* = -0.860). Again, contrary to expectation, this split indicated that the items requiring examinees to recognize a fact or to understand a procedure, an action, or the authors' opinions were more difficult than the item requiring examinees to compare or contrast, to generate a theme, or to apply to the real world.

On the right side of the tree, the 11 items within Node 3 were split into two nodes based on R4 (Knowledge of Discourse Structure Required). The seven items coded as not or somewhat requiring knowledge of discourse structure were assigned to the left node (*Item Mean* = 1.461), and the four items coded as critically requiring knowledge of discourse structure were assigned to the right node (*Item Mean* = 0.552). Contrary to expectation, this split indicated that the items that did not require or somewhat required knowledge of discourse structure were more difficult than the items that critically required knowledge of discourse structure. Subsequently, the items within Node 5 were further split into two nodes based on R1 (Word Recognition Required). The three items coded as not or somewhat requiring the process of R1 were assigned to the left node (*Item Mean* = 1.290), and the four items coded as critically requiring the process of R1 were assigned to the right node (*Item Mean* = 1.590). The items that critically required recognizing words with advanced phonological or orthographical knowledge or identifying the meaning of an unknown word with few context cues were more difficult than the items that did not or somewhat required such processes. Next, the items within Node 9 were further split into two terminal nodes based on T2 (Type of Match). The two items coded as having a literal or synonymous match between question stem and text were assigned to the left terminal node (*Item Mean* = 1.380), and the single item coded as having no match between question stem and text was assigned to the right terminal node (*Item Mean* = 1.110). Contrary to expectation, this split indicated that the items that required the match between question and text were more difficult than the item that did not. The items within Node 10 were evenly split into two terminal nodes based on R2 (Vocabulary Knowledge Required). The two items coded as not or somewhat requiring the process of R2 were assigned to the left terminal node (*Item Mean* = 1.505), and the two items coded as critically requiring the process of R2 were assigned to the right terminal node (*Item Mean* = 1.675). This split indicated that the items that critically required knowledge of infrequently used and/or specialized vocabulary were more difficult than the items that did not or somewhat required such knowledge.

Returning to Node 6, the four items in Node 6 were split into two nodes based on T2 (Type of Match). The single item coded as having literal match between question stem and text was assigned to the left node (*Item Mean* = 1.110). The three items coded as having a synonymous match or no match between question stem and text were assigned to right node (*Item Mean* = 0.367). Again, contrary to expectation, this split indicated that the item requiring a literal match between question and text was more difficult than the items requiring a synonymous match or no match between question and text. Due to the lack of child nodes, the left node became a terminal node. Lastly, the items with Node 11 were split into two terminal nodes based on R6 (Inference Required). The single item coded as requiring no or low-level inference was assigned to the left terminal node (*Item Mean* = 0.090), and the two items coded as requiring high-level inference were assigned to the right terminal node (*Item Mean* = 0.505). This split indicated that the items requiring high-level inference were more difficult than the item requiring no or low-level inference.

The final tree for the 10-component cognitive model using the Form F data included nine components, produced 12 terminal nodes, and involved four levels of splits. The first split at the root node explained a substantial amount (67.3%) of the variance in item difficulty. The two splits at the second level increased the amount of variance explained to 93.2%. The three splits at the third level increased the variance explained to 97.3%. The five splits at the last level increased the portion of variance explained to 99.3%.

Test of the 11-component cognitive model (Form F). The TBR analysis of the cognitive model with 11 components using the Form F data produced an initial tree that included ten predictors: T3, R6, R4, R3, R1, T2, R7, T1, R5 and T5; R2 was the only predictor that was not included in the tree. This tree produced 14 terminal nodes and accounted for 99.5% of the variance in item difficulty (see Appendix M). However, when the four terminal nodes produced by the split based on T1 and R5 were pruned, the resulting tree had 12 terminal nodes and explained 99.4% of the variance. Given the small change in variance explained and the desire for parsimony, the latter tree containing eight predictors (T3, R6, R4, R3, R1, T2, R7, and T5) was selected as the final tree for the 11-component cognitive model using the Form F data (see Figure 13).

As seen in Figure 13, the first variable entering the tree was, again, T3 (Number of Plausible Distractors), which separated the 20 items into two nodes. The left node contained 9 items coded as having zero or one plausible distractor (*Item Mean* = -0.442). The right node contained 11 items coded as having two or three plausible distractors (*Item Mean* = 1.131). This split again indicated that items with more than one plausible

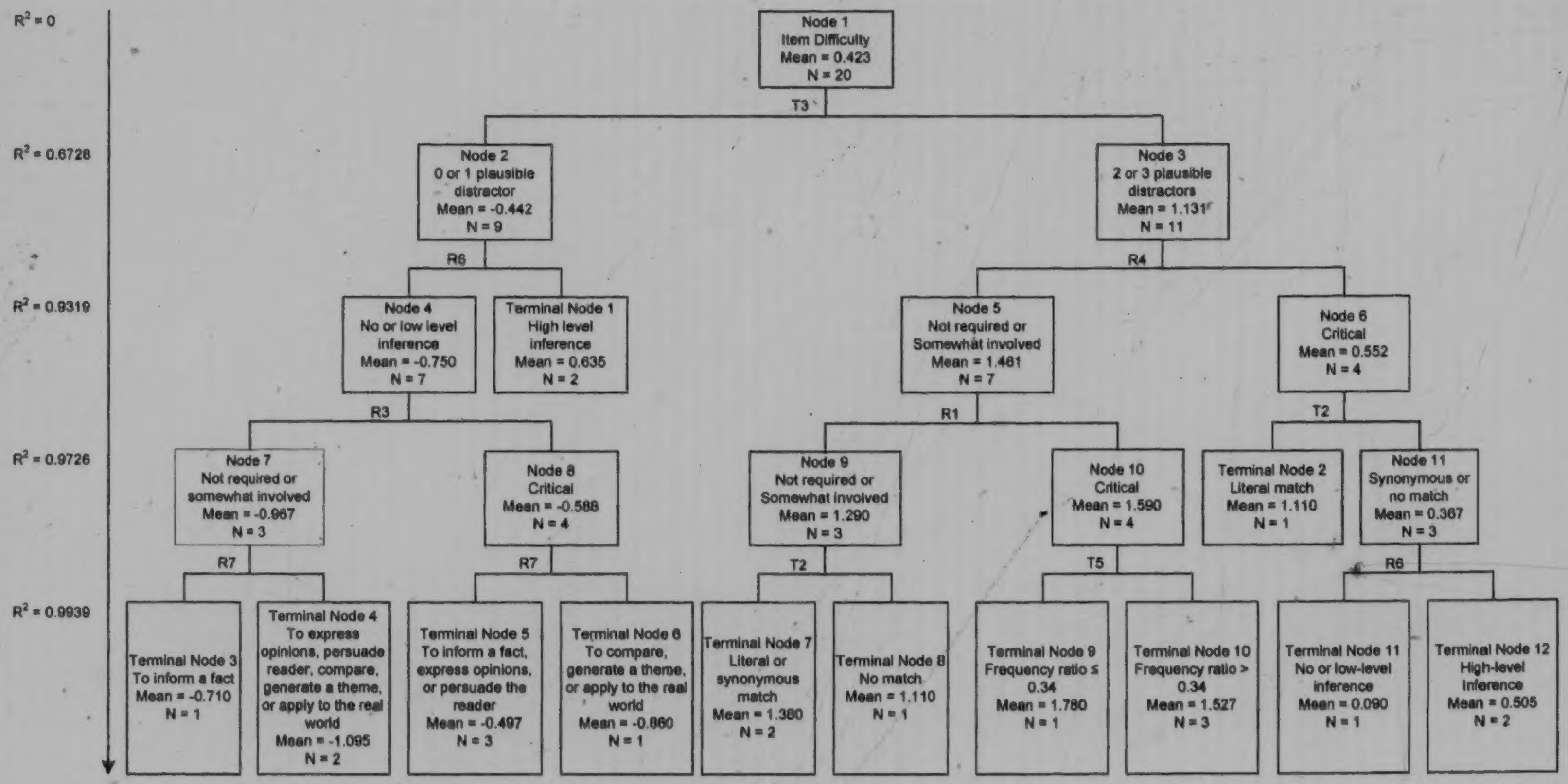

$R^2 = 0$
$R^2 = 0.6728$
$R^2 = 0.9319$
$R^2 = 0.9726$
$R^2 = 0.9939$
Node 1 Item Difficulty Mean = 0.423 N = 20
T3
Node 2 0 or 1 plausible distractor Mean = -0.442 N = 9
Node 3 2 or 3 plausible distractors Mean = 1.131 N = 11
R6
R4
Node 4 No or low level inference Mean = -0.750 N = 7
Terminal Node 1 High level inference Mean = 0.635 N = 2
Node 5 Not required or Somewhat involved Mean = 1.461 N = 7
Node 6 Critical Mean = 0.552 N = 4
R3
R1
T2
Node 7 Not required or somewhat involved Mean = -0.967 N = 3
Node 8 Critical Mean = -0.586 N = 4
Node 9 Not required or Somewhat involved Mean = 1.290 N = 3
Node 10 Critical Mean = 1.590 N = 4
Terminal Node 2 Literal match Mean = 1.110 N = 1
Node 11 Synonymous or no match Mean = 0.367 N = 3
R7
R7
T2
T5
R6
Terminal Node 3 To inform a fact Mean = -0.710 N = 1
Terminal Node 4 To express opinions, persuade reader, compare, generate a theme, or apply to the real world Mean = -1.095 N = 2
Terminal Node 5 To inform a fact, express opinions, or persuade the reader Mean = -0.497 N = 3
Terminal Node 6 To compare, generate a theme, or apply to the real world Mean = -0.860 N = 1
Terminal Node 7 Literal or synonymous match Mean = 1.380 N = 2
Terminal Node 8 No match Mean = 1.110 N = 1
Terminal Node 9 Frequency ratio ≤ 0.34 Mean = 1.780 N = 1
Terminal Node 10 Frequency ratio > 0.34 Mean = 1.527 N = 3
Terminal Node 11 No or low-level inference Mean = 0.090 N = 1
Terminal Node 12 High-level Inference Mean = 0.505 N = 2

Figure 13. Tree diagram for the cognitive model with 11 components (Form F).

distractor were more difficult than items with zero or one plausible distractor.

The items within Node 2 were then split into two nodes based on R6 (Inference Required). The left node contained seven items coded as requiring no or low- level inference (*Item Mean* = -0.750) and the right node contained two items coded as requiring high-level inference (*Item Mean* = 0.635). This split indicated that item requiring high-level inferences were more difficult than items requiring no or low-level inferences. As the right node did not have any child nodes, it became a terminal node. Next, the seven items within Node 4 were further split into two nodes based on R3 (Syntactic Knowledge Required). The left node contained three items coded as not or somewhat requiring syntactic knowledge (*Item Mean* = -0.967), and the right node contained four items coded as critically requiring syntactic knowledge (*Item Mean* = -0.588). This split indicated that the items that required knowledge of complex infrequently used syntax were more difficult than the items that did not. The three items within Node 7 were further split into two terminal nodes based on R7 (Purpose of Information). The single item coded as to inform a fact was assigned to the left terminal node (*Item Mean* = -0.710). The two items coded as to state a procedure or action, to express the authors' opinions, to persuade the reader, to compare or contrast, to generate a theme, or to apply to the real world were assigned to the right terminal node (*Item Mean* = -1.095). Contrary to expectation, this split indicated that the item requiring examinees to recognize a fact was more difficult than the items requiring examinees to understand a procedure, an action, or the authors' opinions, to compare or contrast, to generate a theme, or to apply to the real world. Lastly, the items within Node 8 were also split into two terminal nodes based on R7 (Purpose of Information). The three items coded as to inform a fact, state a procedure or action, express opinions, or to persuade the reader were assigned to the left node (*Item Mean* = -0.497), and the single item coded as to compare or contrast, generate a theme, or to apply to the real world was assigned to the right node (*Item Mean* = -0.860). Again, contrary to expectation, this split indicated that the items requiring examinees to recognize a fact or to understand a procedure, an action, or authors' opinions were more difficult than the item requiring examinees to compare or contrast, to generate a theme, or to apply to the real world.

On the right side of the tree, the 11 items within Node 3 were split into two nodes based on R4 (Knowledge of Discourse Structure Required). The seven items coded as not or somewhat requiring knowledge of discourse structure were assigned to the left node (*Item Mean* = 1.461), and the four items coded as critically requiring knowledge of discourse structure were assigned to the right node (*Item Mean* = 0.552). Contrary to expectation, this split indicated that the items that did not require or somewhat required knowledge of discourse structure were more difficult than the items that critically required knowledge of discourse structure. Subsequently, the items within Node 5 were further split into two nodes based on R1 (Word Recognition Required). The three items coded as not or somewhat requiring the process of R1 were assigned to the left node (*Item Mean* = 1.290), and the four items coded as critically requiring the process of R1 were assigned to the right node (*Item Mean* = 1.590). The items that critically required recognizing words with advanced phonological or orthographical knowledge or identifying the meaning of an unknown word with few context cues were more difficult than the items that did not or somewhat required such processes. Next, the three items within Node 9 were further split into two terminal nodes based on T2 (Type of Match). The two items coded as having a literal or synonymous match between question and text were assigned to the left terminal node (*Item Mean* = 1.380), and the single item coded as having no match between question and text was assigned to the right terminal node (*Item Mean* = 1.110). Contrary to expectation, this split indicated that the items that required the match between question and text were more difficult than the item that did not. The items within Node 10 were also split into two terminal nodes based on T5 (Metacognitive Strategies Used). The single item for which metacognitive strategies were less frequently used (frequency ratios ≤ 0.34) was assigned to the left terminal node (*Item Mean* = 1.780), and the three items for which metacognitive strategies were more frequently used (frequency ratios > 0.34) were assigned to the right terminal node (*Item Mean* = 1.527). This split indicated that the item for which metacognitive strategies were less frequently reported by the students was more difficult than the items for which metacognitive strategies were more frequently reported.

Returning to Node 6, the four items in Node 6 were split into two nodes based on T2 (Type of Match). The single item coded as having a literal match between question and text was assigned to the left node (*Item Mean* = 1.110), and the three items coded as having a synonymous match or no match between question and text were assigned to right node (*Item Mean* = 0.367). Contrary to expectation, this split, again, indicated that the item requiring a literal match between question and text was more difficult than the items requiring a synonymous match or no match between question and text. Due to the lack of child nodes, the left node became a terminal node. Lastly, the three items with Node 11 were also split into two terminal nodes based on R6 (Inference Required). The single item coded as requiring no or low-level inference was assigned to the left terminal node (*Item Mean* = 0.090), and the two items coded as requiring high-level inference were assigned to the right terminal node (*Item Mean* = 0.505). This split indicated that the items requiring high-level inference were more difficult than the item requiring no or low-level inference.

The final tree for the 11-component cognitive model using the Form F data included eight components, produced 12 terminal nodes, and involved four levels of splits. The first split explained a substantial amount (67.3%) of the variance in item difficulty. The two splits at the second level increased the amount of variance explained to 93.2%. The three splits at the third level increased the variance explained to 97.3%. The five splits at the last level increased the portion of variance explained to 99.4%.

A comparison of the 10- and 11-component cognitive model (Form F). A comparison of the final tree solutions for the 10- and 11-component cognitive model using the Form F data, as shown in Figures 12 and 13, revealed three points. First, for the 10-component cognitive model, the final tree included nine components in the cognitive model: T3, R6, R4, R3, R1, T2, T1, R7, and R2; R5 was the only component that was not included in the final tree. However, for the 11-component cognitive model, the final tree included eight of the 11 components in the cognitive model; three components – R5, R2, and T1 – were not included in the final tree. That R5 did not appear in the Form F trees for both the 10- and 11-component cognitive model was likely due to its high correlation with R4 ($r = 0.750, p < 0.05$) and moderately high correlation with R6 ($r = 0.631, p < 0.05$). In other words, part of the information contained in R5 (Synthesis Required) may be captured in the variables R4 (Knowledge of Text Structure Required) and R6 (Inference Required). In contrast, for Form E, R5 had moderately high correlation with R4 only ($r = 0.685, p < 0.05$). The correlation between R5 and the other variables were weak. The correlation matrix for Forms E and F are provided in Appendix N and Appendix O, respectively.

Second, the final tree solutions for the 10- and 11-component cognitive model using the Form F data were similar. After pruning, the final tree solutions for both cognitive models produced four levels of split. Moreover, the root node, the first three levels of splits, and the splits of the items in Node 8, Node 9, and Node 11 based on R7, T2, and R6, respectively, at the last level of split were in the same positions and produced exactly the same results for both cognitive models. However, the items within Node 7 and Node 10 were split, respectively, by T1 (Location of Information) and R2 (Vocabulary Knowledge Required) in the 10-component model, and respectively by R7 (Purpose of Information) and T5 (Metacognitive Strategies Used) in the 11-component model.

Third, the 10-component item difficulty model explained 99.3% of the variance in item difficulty. The 11-component item difficulty model explained only 0.1% more

(99.4%) of the variance than the 10-component model. Hence, the inclusion of T5 in the cognitive model did not lead to a substantial increase in the variance explained in item difficulty. Given the statistical principle of parsimony (Kerlinger, 1979), it appeared that the cognitive model without T5 is a better model. Overall, the TBR results for Form F supported the 10-component cognitive model.

A comparison of the 10- and 11-component model across the test forms. The results from the two sets of TBR analyses for the two test forms are summarized in Table 17. The table displays the components included and not included in each final tree, total number of splits, whether pruning was conducted or not, total number of terminal nodes, and the amount of variance explained by the final tree. An examination of Table 17 reveals that for both 10- and 11-component models, the final trees had comparable number of terminal nodes across the two forms. Further comparison of the final trees for the 10- and 11-component cognitive models across the two forms (Figure 10 *vs.* Figure 12; Figure 11 *vs.* Figure 13) reveals that T3 (Number of Plausible Distractors) was the first variable that was included in both the Form E and the Form F final trees, and that the first split of the Form F final trees explained much more variance than the first split of the Form E final trees (67.3% *vs.* 34.3%). Moreover, for both 10- and 11-component models, all variables but T3 entered the Form E and Form F trees in different orders. This was expected given that (1) the degree to which each process was called for by the items in the two forms were not strictly parallel, (2) the distribution of item difficulty was different across the two forms, (3) the TBR tended to select variables that can produce

Table 17

The Final Trees for the Two Cognitive Models and Two Forms[1]

Cognitive Model	Form	Entered the Tree	Not Entered the Tree	Levels of Split	Pruning Occurred	# of Terminal Nodes	R^2
10 component	E	T3, R3, R1, R5, R4, T1, R2, T2, R7, R6	--	5	No	13	97.9%
	F	T3, R6, R4, R3, R1, T2, T1, R7, R2	R5	4	Yes	12	99.3%
11 component	E	T3, R3, R1, R5, R4, T1, R2, R7, T5	R6, T2	4	Yes	11	91.9%
	F	T3, R6, R4, R3, R1, T2, R7, T5	R5, R2, T1	4	Yes	12	99.4%

Note. [1] Abbreviations used in the table: T3 = Number of Plausible Distractors; R3 = Syntactic Knowledge Required; R1 = Word Recognition Required; T1 = Location of Information; R2 = Vocabulary Knowledge Required; R4 = Knowledge of Discourse Structure Required; T5 = Metacognitive Strategies.

more splits in the tree-growing process, and (4) in the case where two or more different predictors produced an equal improvement statistic for a given split, the selection among these variables to split this node is simply based on the order of variables in the TBR analysis (AnswerTree User's Guide, 2002).

Table 17 also reveals the differences between the 10- and 11-component cognitive model in light of the similarities and discrepancies of TBR results across the two forms. For the 10-component cognitive model, given that there was no need to prune, the final tree for Form E had five levels of split. However, as the result of pruning, the final tree for Form F had four levels of split. For the 11-component cognitive model, following pruning, the final tree for both Form E and Form F had four levels of split.

However, compared with the 10-component cognitive model, the TBR analyses of the 11-component cognitive model produced largely divergent results across the two test forms. For the 10-component cognitive model, the final tree for Form E and the final tree for Form F had nine components overlapped, and the two final trees explained comparable amounts of variance (97.9% *vs.* 99.3%). For the 11-component cognitive model, the final tree for Form E and the final tree for Form F had only six components overlapped, and there was a relatively large difference between the total amount of variance explained by the two final trees (91.9% *vs.* 99.4%). Overall, the comparison of the TBR results for the 10- and 11-component cognitive model across the two forms provided further evidence in support of the 10-component model.

In sum, the 10-component cognitive model was superior to the 11-component cognitive model mainly for two reasons. First, for both forms, the final tree for the 11-component cognitive model included fewer components in the tree and did not explain substantially more variance than the 10-component cognitive model. Second, the 10-component cognitive model had more variables that overlapped between the Form E and Form F final trees and explained comparable total amounts of variance across the two forms. Therefore, the 10-component cognitive model was considered more effective and parsimonious, and was thus retained as the final cognitive model for the MELAB reading test items.

As noted above, the TBR analyses produced a total of seven unexpected results of splitting, indicating that cognitively more demanding items were easier than cognitively less demanding items. These splits were, respectively, based on R5, R1, and R4 in the final tree for Form E and on R4, T1, T2, and R7 in the final tree for Form F (see Figures 10 and 12). One possible explanation for these ambiguous findings may be the way in which the items were initially coded. For example, the result of the split based on R5 indicated that the items that required synthesis were easier than the items that did not. An item was initially coded as "No synthesis" if the information requested by the item was contained within a single sentence, "Low-level synthesis" if the requested information was contained within multiple adjacent sentences, and "High-level synthesis" if the requested information was contained within multiple nonadjacent sentences or diffused across the passage. A problem with the present coding of R5 may be that the level of synthesis required to correctly answer an item might not be adequately evaluated simply based on the distribution of the requested information. Rather, synthesis involves complex thinking processes. The objective scoring used to code the items did not fully capture the complexity of actual thinking of the students, thus leading to inaccurate prediction of the item difficulty.

Item ambiguity is another possible explanation for the unexpected findings. In the Form F final tree, the split based on R4 indicated that the four items (i.e., items 6, 9, 14, and 19) that critically required knowledge of discourse structure were easier than the seven items (i.e., items 4, 12, 13, 15, 17, 18, and 20) that did not require or somewhat required this knowledge. Examination of the seven items assigned to the left node revealed that they were the seven most difficult items on Form F, with item difficulty estimates ranged from 1.11 to 1.78 (see Table 16). Further examination of the students' verbal reports revealed that the poor quality of the items 12, 13, and 15 made them extremely difficult to understand, and thus leading to spuriously high difficulty, despite the cognitive processes assessed by these items. For example, student FS3 reported,

> This item (Item 12) is not well designed. I really think none of the options is correct. C appears to be the closest to the correct response, but it is not strictly logical. I expect the correct answer is "have different needs at the *same* time". Option C is "have different needs at different time". It is not logically correct, because "different needs at different time" involves the possibility of having the same needs at the same time. The passage does not mention "different needs at different time". In fact, "same time, different needs" is the key to "competitive exclusion". It does not matter whether they compete or not at different time. The key point of the passage is that they have different needs at the same time… This item (Item 13) does not have a sound answer. Monoculture means to plant the same type of crops in the same field, while the only seemingly correct response "Standard method" refers not only to plant the same type of crops in the same field, but also to plant the same crops each season. Hence, "Standard method" has a broader meaning than "Monoculture". It is not a

> good answer, but I have to choose it…. This item (Item 15) is very confusing. The question asks the reasons for overyield. But none of the options explains the reason.

In addition to reading test items, the difficulty of the reading passages may have affected item difficulty. The difficulty of test passages is affected by a variety of factors, including text topic and content, text type and genre, text organization, and linguistic variables such as syntactic complexity (Alderson, 2000). The items referenced to easier passages tend to be easier than the items referenced to more difficult passages, despite the cognitive processes measured by the items (see, also, Alderson, 2000; Huff, 2003). This might be a possible explanation for the unexpected result of splitting produced by R1 in the Form E final tree (see Figure 10). This split indicated that the six items that critically required the process of word recognition (Items 2, 3, 5, 6, 8, and 16) were easier than the six items that did not require or somewhat required this process (Items 10, 12, 15, 17, 19, and 20). An interesting finding was that other than Item 10, the other five items in the left node were based on Passages 3 and 4, and that other than Item 16, the other five items in the right node were based on Passages 1 and 2. It is likely that Passages 3 and 4 were more difficult than Passages 1 and 2. Hence, while R1 was not critically required to correctly answer these items, they were still more difficult than items that required R1 but were based on relatively easier passages. Due to the small number of items, only exploratory TBR analyses were conducted in the present study. Exploratory analysis simply considers the maximization of deviance statistically. However, the single process that statistically maximizes the deviance at each node does not necessarily represent the single most important process used by the students to arrive at the correct response. Confirmatory analyses using the TBR may be conducted in future when larger number of items is available to improve the model interpretability.

---

[1] The equation used to calculate the *MSR* is $MSR = \frac{\sum_{i=1}^{n} [X_i(\theta) - Y_i(\theta)]^2}{n-1}$, where n represents the number of score points on the theta scale and in the context of comparison of the TCCs of Form E and Form F, $X_i(\theta)$ is the expected total score at θ for Form E, and $Y_i(\theta)$ is the expected total score at θ for Form F. The null and alternative hypothesis tested for *MSR* are H_0: $MSR = 0$; H_1: $MSR > 0$.

The value of the MSR was compared to the critical value in a chi-square distribution with n-1 degrees of freedom to test whether the MSR is statistically different from 0 or not. If the MSR is statistically different from 0, then the two TCCs are said to be statistically different from one another. Similarly, if the MSR is not statistically different from 0 then the two TCCs are said to be statistically similar to each other. Thirty score points with equal intervals were selected from the theta scale in this study.

CHAPTER 7: SUMMARY, CONCLUSION, AND RECOMMENDATIONS

Given the sequential nature of this study, the methods for each stage and the findings obtained are summarized together. The limitations of the study are then discussed, followed by the conclusion drawn in light of the limitations. Lastly, practical implications of the findings in this study are addressed, followed by directions for future research.

Summary of Methods and Discussion of the Findings

In the present study, a cognitive processing model for the MELAB reading test items was developed and tested through four stages. First, informed by substantive theories of L2 reading processes and ability constructs and research into L2 reading test and item performance, an initial cognitive model was hypothesized to underlie the MELAB reading item performance. Then, this model was empirically tested through cognitive analysis of the test items by raters, students' verbal protocols of the actual processes they used to arrive at the correct responses, and item difficulty modeling with the TBR.

Development of the Initial Cognitive Model

A review of the pertinent literature on the processes associated with reading and reading test taking by L2 learners suggested that seven reading processes, four test-management processes, and three testwise processes should be included in the initial cognitive model underlying the L2 reading test item performance. These processes included: Word recognition (R1), Using vocabulary knowledge (R2), Using syntactic knowledge (R3), Using knowledge of discourse structure (R4), Synthesizing (R5), Drawing inferences (R6), Using pragmatic knowledge (R7), Locating specific details in text (T1), Matching question to text (T2), Evaluating alternative choices (T3), Using topical knowledge (T4), Using clues in other items (TW1), Guessing (TW2), and Using surface features of answer choices (TW3).

Operational definitions for each of the 14 cognitive processes were then developed in the context of responding to multiple-choice reading test items. Next, potential sources of processing difficulty were identified for each process. These sources focused on different aspects of reading test items: processing of information, location of requested information, familiarity of passage topic, and characteristics of item stems and options. Lastly, scoring scales were defined to score the cognitive complexity of the processes on the basis of theoretical and empirical relationships informed by previous research (Bachman et al., 1996; Jamieson et al., 2000; Rupp et al., 2001; Sheehan & Ginther, 2001).

Efforts were made to cover a wide range of processes thought to be involved when answering multiple-choice reading test items. Consistent with the emphasis of reader-text interaction and the importance of test method in current literature of L2 reading assessments (e.g., Alderson, 2000; Bachman & Palmer, 1996), the variables that affect the difficulty of reading test items were defined, with both reader and task (text and item) characteristics taken into consideration. Nevertheless, the set of 14 cognitive processes in the initial cognitive model were not meant to be a definitive set. Some processes may have been overlooked and certain definitions of the processes may not be clear. Hence, the initial model was refined and validated using a three-stage procedure.

Cognitive Analysis of the Items by Raters

Three raters familiar with the L2 reading processes and the population of Mandarin-speaking MELAB test takers separately identified which of the 14 cognitive processes in the initial model were necessary to correctly answer the items included in Forms E and F, and to suggest any additional processes required to correctly answer the items. The raters were provided with group training and a rating instrument.

G-theory analysis of the ratings by the three raters suggested a high level of agreement among the three raters' item ratings (*G*-coefficient = 0.90). It appears that the use of a pilot study, rater training, a clearly-defined rating instrument, extensive discussion, exemplification of item coding, and standardized rating procedure in the present study contributed to the agreement among the raters. This level of agreement is consistent with other researchers (e.g., Bachman et al., 1996; Carr, 2003; Xi, 2003), who used well-designed rating instruments, exemplification, and re-categorization of the cognitive demands of the test items.

The final set of consensus ratings, which was generated through a thorough discussion of the reading items included in the two forms, revealed several points. First, the seven reading processes and the first three test-management processes in the initial cognitive model were identified by the raters as required to correctly answer the items in the two forms: R1, R2, R3, R4, R5, R6, R7, T1, T2, and T3. However, the raters did not identify the fourth test-management process and the three testwise processes. Second, the raters did not identify any additional processes that would be needed to correctly answer the reading items included in the two forms. However, failure to identify additional processes may also be caused by presenting the processes to the raters. While the raters were encouraged to identify any processes not included in the initial model, their perspectives might have been limited by the available processing components. Third, the raters confirmed that the scoring scales reflected discernible levels of performance for the processes considered. Fourth, the raters indicated that simultaneous use of multiple cognitive processes was required to correctly answer reading test items. This finding supports the view of some researchers that reading can be divided into specific cognitive elements, which are used in combinations to correctly interpret what has been read or to correctly answer reading test items (e.g., Gorin, 2002; Grabe, 1991). Lastly, a comparison of the cognitively-based item features coded by the raters across the two test forms revealed that the complexity and demands of the cognitive processes required by the items included in the two test forms were different, which suggested that the distributions of item difficulty might be different across the two forms.

The rater results linked the initial cognitive model to the MELAB reading items, corroborated the scoring scales defined in the initial model, and provided evidence in support of a revised cognitive model. However, as Alderson (2005a) and Leighton and Gierl (2005) pointed out, judges may process test tasks differently from the target examinees. Hence, examinees' actual processes underlying the correct responses were investigated to determine the processes actually used by examinees when they responded to the MELAB reading test items.

Student Protocols and Analysis

Verbal report data were collected from 16 Mandarin-speaking undergraduate and graduate students as they answered the reading items on Form E or Form F. Both concurrent and retrospective data were collected. Each interview was completed in two sessions scheduled within a week, with two reading passages and their associated items administered in each session. The researcher coded the verbal data first using the coding scheme based on the processing components in the initial cognitive model, and then using the finer coding scheme based on the scoring scales used by the raters. A trained coder independently coded 37.5% of the verbal data using the same coding schemes as those used by the researcher. The codes assigned by the researcher were compared to those assigned by the independent coder. The percentage of total agreement between the researcher and the independent coder was 82.8%, which indicated that the cognitive processes segments were consistently coded.

The verbal results supported the inclusion of R1, R3, R4, R5, R6, R7, T1, T2, and T3 and warranted the elimination of T4, TW1, TW2, and TW3. However, the verbal results did not support R2 (Using vocabulary knowledge) and raised an additional component T5. Leighton (2004) reminded us that students' verbal reports were sensitive to task demands, and that they were difficult to obtain when "the task used to elicit the reports was exceedingly difficult or called upon automatic processes" (p. 12). R2, which was not found in the students' verbal reports, likely was an automatic process for the students who were quite proficient in English. In such cases, as suggested by Leighton (2004), raters' perspectives of the cognitive processes required to correctly answer the items would reveal the automatic processes that could not be verbalized by the students. Given that R2 was identified by the raters as needed to correctly answer the items and that prior research in L2 reading revealed a close relationship between R2 and reading comprehension (e.g., Bachman et al., 1996; Carr, 2003; Johnston, 1983), R2 was retained in the revised cognitive model.

A new category, T5 (Using metacognitive strategies) emerged from the verbal report data. T5 was, respectively, the fourth and the third most frequently reported process for Forms E and F. Using metacognitive strategies while answering the reading test items is included in Bachman's (1990) CLA framework. Further, a number of research studies have found that good readers are more effective in using metacognitive strategies and more capable of describing the use of such strategies (e.g., Block, 1992; Grabe, 1991; Johnston, 1983; Phakiti, 2003). Consistent with the findings in these studies, the proficient ESL students in the present study were good at using and describing the metacognitive strategies they used while answering the reading test items. However, since T5 was not identified by the raters, the inclusion of T5 was undetermined. Hence, two revised cognitive models were retained at this point of the study. The first one contained the first 10 components in the initial cognitive model, and the second contained these 10 components plus T5.

In agreement with the rater findings, the verbal results revealed that different processes were involved in responding to different items and simultaneous use of multiple processes was needed to correctly answer an item. However, unlike the rater findings, which revealed that the scoring guides for different cognitive processes were valid and useable, the verbal data were not sensitive enough to discern different levels of performance for processes such as R3 (Syntactic knowledge required). This finding lends support to Leighton's (2004) comments on the advantages of using both raters' analysis of the cognitive processes required to correctly answer test items and analysis of students' think-aloud protocols to justify the cognitive processes underlying test item performance.

Taken together, the rater and verbal data revealed processes noted in other studies of the processes used by L2 readers during taking academic reading tests (e.g., Abbott, 2005; Alderson, 1990b; Enright et al., 2000; Hudson, 1996; Jamieson et al., 2000). The two sets of findings support the testing of two revised cognitive models with the TBR analyses.

Item Difficulty Modelling using Tree-Based Regression

Two sets of TBR analyses were conducted, one using the Form E data and the other using the Form F data to test the cognitive model with 10 and 11 components, respectively. The TBR results supported the cognitive model with 10 components. For both test forms, the final trees for the 11-component model included fewer components and did not explain substantially more variance than the final trees for the 10-component model. Further, the 10-component model had more variables that overlapped between the Form E and Form F final trees and explained comparable total amounts of variance across the two forms. Therefore, the 10-component model was considered more effective and parsimonious, and was thus retained as the final cognitive model for the MELAB reading test items. These 10 processing components have been included in other models of the L2 reading processes, albeit not together (e.g., Alderson, 2000; Koda, 2005), and have been revealed as important predictors in studies of the factors affecting reading item or test performance (e.g., Drum et al., 1981; Freedle & Kostin, 1993; Rupp et al., 2001; Sheehan & Ginther, 2001).

However, the final trees for Forms E and F were not identical: (1) while T3 led to the first split, the first split of the Form F tree explained much more variance than that of the Form E tree, (2) except for T3, the remaining variables entered the two trees in different orders, and (3) R5 entered the Form E tree only. The divergent tree structures were likely due to two reasons. First, as indicated by the rater and verbal results, the degree of cognitive processes called for by the reading items was different across the two test forms. Second, the psychometric characteristics of the two forms revealed that while the statistics such as independent samples *t*-test and *MSR* supported parallelism of the two test forms, the distributions of item difficulty were different across the two forms. Form F contained both more difficult and more easy items than Form E. These findings speak for the complexity of item construction and remind us that caution needs to be exercised when interpreting reading performance on different test forms. To better understand the validity of test scores, not only should the statistical properties of a test, such as mean, variance, and reliability, be examined, but also the cognitive processes elicited by the items need to be analyzed to provide substantive evidence regarding the nature of constructs assessed by the test. As Gorin (2002) noted, both statistical and substantive analyses of the items are required to effectively construct and evaluate reading test items. More importantly, test developers need to carefully consider the passages used and the nature of the items included at the test construction stage, and to construct tests based on predetermined cognitive processes defined from a cognitive model.

The final trees were relevant to the theoretical constructs of the MELAB reading and explained substantial amounts of the variance in item difficulty (97.9% for Form E and 99.3% for Form F). Huff (2003) developed a tree-based item difficulty model for the new TOEFL reading items. Using already-existing item codes, which were not specifically related to reading item difficulty as predictors, her final tree explained only 56.0% of the item variance. Rupp et al. (2001) modeled item difficulty for a L2 reading test using the TBR. Using features of text, item, and text-item interaction informed by prior studies of reading item difficulty as predictors, their final tree explained only 50.0% of the item variance. However, Sheenhan's (1997) tree-based item difficulty model developed for the items included in the SAT Verbal Reasoning Test explained about 80% of the variance in item difficulty. In her study, the test items were coded using the scoring scales based on the four cognitive processes described in Kirsch and Mosenthal (1990): (1) recognizing vocabulary in context, that is, determining the meaning of the referenced word or phrase by analyzing the surrounding text; (2) understanding of the points explicitly stated in the text; (3) drawing inferences about the author's purpose, assumptions, attitude, or rhetorical strategy; and (4) determining which of several alternative options are best supported by the information in the text. Subsequently, Sheehan and Ginther (2001) developed a tree-based item difficulty model for the Main Idea reading items on the *TOEFL 2000*. Coding the items using three variables describing cognitive processing features of the items (i.e., Correspondence between Correct Response and Text, Location of Relevant Information, and Elaboration of Information), their final tree model explained ever more (87.0%) variance in item difficulty.

In the present study, finely-tuned scorning scales based on 14 cognitive processes were used that were parallel to those used in Sheehan's (1997) and Sheehan and Ginther's (2001) studies. It was not surprising that the final tree models developed in the present study accounted for larger item variance than that was explained in Sheehan's (1997) and Sheehan and Ginther's (2001) studies. The difference between the explained variance in the previous TBR studies and the present study is likely attributable to the use of a more thorough procedure to develop and validate the cognitive processes considered in the TBR analyses conducted in the present study.

Limitations

The verbal report data were collected from a group of Mandarin-speaking students enrolled in a university program. The English capability of these students was well advanced. As mentioned earlier, for the students with high English language proficiency, certain processes such as R2 (Using vocabulary knowledge) were likely automatic and thus could not be identified through verbal description. Examinees with different language backgrounds and proficiency levels likely have mastered different aspects of reading ability to different degrees and likely use different cognitive processes to arrive at their answers (Huff, 2003; Kasai, 1997). As item difficulty is affected by the interaction between examinee and test task (Bachman, 2002), item difficulty is likely to vary across different language groups and proficiency levels. Verbal protocols from students with other first non-English languages and at average and low proficiency levels may reveal more information regarding the cognitive processes required to arrive at the correct responses and increase the degree of correspondence among different sources of data.

The TBR is a tool that can be used to examine how underlying cognitive processes affect item difficulty. However, the findings from the TBR analyses need to be interpreted with caution. The TBR algorithm is designed to select variables at each stage of the tree that account for the greatest addition to explained variance and will continue to do so until the additional amount of explained variance is negligible. At the present time, there is no agreed upon rule regarding what constitutes negligible variance. Moreover, when two or more different predictors produce an equal improvement in explained variance for a given split, the selection among these variables is simply based on the initial order of variables in the TBR analysis (AnswerTree User's Guide, 2002; Ewing & Huff, 2004). Hence, conclusions drawn from the tree structures need to be tempered. In the present study, two forms were considered. While final trees differed between the two forms, different distributions of item difficulties provided an explanation for this difference.

Conclusion

Despite the limitations discussed above, the triangulation of the three sources of evidence collected in this dissertation support the conclusion that seven reading processes

– R1 (Word recognition), R2 (Using vocabulary knowledge), R3 (Using syntactic knowledge), R4 (Using knowledge of discourse structure), R5 (Synthesizing), R6 (Drawing inferences), and R7 (Using pragmatic knowledge) – and three test-management processes –T1 (Locating specific details in text), T2 (Matching question to text), and T3 (Evaluating alternative choices) – underlie successful performance on the MELAB reading test items. Consistent with prior studies using the triangulation approach to determining the test or item performance (e.g., Abbott, 2005; Anderson et al., 1991; Leighton, 2004), this study demonstrated the value of using multiple sources of data to evaluate the performance of a cognitive model, and successfully demonstrated the union of cognitive psychology and assessment in the field of second/foreign language testing.

Implications for Educational Practice

The findings in this study have implications for curriculum and instruction, test development, and diagnostic feedback. First, the findings have implications for ESL/EFL classroom teachers concerning how reading strategies should be taught so that the reading strategies could work best for their students. In this study, both rater and verbal results indicated that the simultaneous use of multiple cognitive processes is needed to understand the texts and correctly answer the reading test items. In light of this finding, as noted by Grabe (2004), reading strategies may be taught through a combined-strategies instructional approach rather than taught independently of one another.

Second, the findings provide guidelines for the MELAB test takers concerning how to prepare for the MELAB reading tests effectively. The finding that the cognitive model underlying the MELAB reading item performance did not include the four construct-irrelevant processes – T4 (Using topical knowledge), TW1 (Using clues of other items), TW2 (Guessing), and TW3 (Using surface feature of answer choices) – suggested that stressing these processes would not be effective. Rather than focusing on learning and practicing how to become testwise, the MELAB test takers should focus on improving their English language knowledge and strategic competence.

Third, the findings can be used to guide test developers to design cognitively-based reading test items (Gorin, 2002; Embretson, 1999). While the TBR analyses produced somewhat divergent results across the two test forms, the pattern of agreement between the final trees for Forms E and F sheds some light on which of the construct-relevant item features most likely affected the MELAB reading item difficulties. For example, the final trees for Forms E and F both suggested that items with more plausible distractors were more difficult than items with less or none plausible distractors, and that items requiring high-level inference were more difficult than items requiring no or low-level inference. Further, the resulting tree models suggested some mechanisms for ordering the items based on the relationship between cognitively-based item features and item difficulty. For example, an item bearing the following features would be an easy item: it does not have plausible distractors and requires simple syntactic knowledge. In contrast, an item bearing the following features would be a difficult one: it has more than one plausible distractor, somewhat requires recognizing the meaning of unknown words using context clues, and requires information located in the entire passage. When item specifications are written in terms of cognitive processes to be assessed by each item, more direct assessment of what the students know and what they can do may be achieved.

Fourth, the findings from the TBR analyses can be used to provide detailed diagnostic feedback to the examinees and lay a foundation for the MELAB as a diagnostic measure. The tree models obtained in this study produced homogeneous clusters of items that measure similar cognitive processes and have similar item difficulty. By summarizing examinee performance against these item clusters, meaningful diagnostic score reports can be generated to suggest particular knowledge or skills that the examinee needs to work on. For example, if an examinee responds incorrectly to most of the items within a cluster measuring knowledge of complex syntax, conclusions can be drawn that this examinee has trouble understanding the sentences with complex syntax in academic texts and needs more work to gain such knowledge. If an examinee correctly answers the cluster of items measuring the competence of recognizing infrequently used or specialized vocabulary in academic texts, then conclusions can be drawn that this examinee can correctly identify vocabulary in academic texts, which is required for academic studies in an English-speaking setting.

Directions for Future Research

Although empirical evidence has been found in support of the 10-component cognitive model for successful performance on the MELAB reading forms, the model warrants further research so that other large-scale language testing programs designed for L2 and large-scale reading tests for L1 could yield more meaningful results that validly reflect examinees' reading ability. In this study, the use of students' verbal protocols provided insights into what the examinees thought and where they attended to while answering the MELAB reading test items. However, the generalization of the findings of this study is constrained by the characteristics of the participants, the data collection procedure, and the data analysis procedure used for this stage of the study. As noted earlier, the verbal protocol participants were Mandarin-speaking students using English in a university setting. Hence, the findings generalize to this population only, which is characterized as Mandarin first language with relatively high English reading proficiency. This stage of the study should be replicated in an examinee population with more heterogeneous English language proficiency and native language backgrounds. The verbal protocols of the students who are about to take the MELAB would also provide a clearer picture of various cognitive processes underlying the MELAB reading test item performance.

The students' verbal protocols were collected to shed light on what cognitive processes were actually used by examinees to correctly answer the MELAB reading test items. A look at the cognitive processes used by the students for the incorrect responses may reveal meaningful information as to what cognitive processes are truly required to solve an item. However, a limitation with the verbal report data is that not all cognitive processes are reported or reported every time when they are used (see, also, Cohen & Upton, 2006). Hence, it is unclear whether the processes not verbalized were indeed not used by the students. Interview strategies are needed to uncover whether or not processes such as R2 are automatic or not, and where the strategies lead the interviewee to produce an answer that is not true or valid.

The use of TBR to model the reading item difficulty can enrich understanding of the relationship between item features and item difficulty, and contribute to the progress of new methods for language testing research. However, the TBR technique has only been applied to a few language studies. Further research focusing on the evaluation of this measurement model is required. Among the items to consider are the stopping rule, the order in which competing variables at a given stage are selected, and whether modeling other item statistics, such as item discrimination, in terms of the cognitive processes involved in item solving may reveal more meaningful information for test developers. Another area of research would be to compare the TBR with other currently available measurement models, such as structural equation modeling, multivariate generalizability theory, and the attribute hierarchy model to examine the extent to which these measurement models adequately describe the response data that is the product of the interaction between reader abilities and test task characteristics.

In addition to the areas of future research mentioned thus far, research in four other directions hold promise. First, the combination of substantive theories of L2 reading and three sources of empirical data provided substantial support for the revised cognitive model containing 10 components. Further cross-validation studies using other MELAB reading test forms are required to determine whether the 10-component cognitive model holds with other test forms, whether T5 needs to be added, and whether the item features identified in this study are stable across different test forms. Second, studies in which test items are first developed using predetermined cognitive processes defined from a cognitive model and then validated using empirical evidence such as that considered in the present study may lead to tests that yield scores that can be more validly interpreted. Third, a cognitive model of the MELAB reading item performance was proposed in this study. Using substantive evidence along with statistical results to generate and validate the cognitive model, this research provides measurement specialists and content experts with rich validity information and characteristics of complex task performance, which will contribute to a better understanding of how students solve test tasks or items. Future studies are required to determine how best to meld the tree-based item difficulty model developed in this study with its cognitive base and with classical and item response model item analyses. Lastly, as mentioned in the earlier chapter, the way of item coding was a likely reason for the ambiguous findings from the TBR analyses. Further research is needed to improve the way in which the items were coded to clarify the role of R5 and to determine why the TBR analyses produced unexpected results of splits. An interesting attempt might be to first more clearly distinguish between two sets of factors that can affect item difficulty: cognitive processes used by examinees and characteristics inherent in the item itself, and then to separately examine the relationship of item difficulties to process ratings and task characteristic ratings, as suggested by Bachman (2002) and executed by Bachman et al. (1996).

References

Abbott, M. L. (2005). *English reading strategies: Differences in Arabic and Mandarin speaker performance on the CLBA reading assessment.* Unpublished PhD dissertation, University of Alberta.

Alderson, J. C. (2005a). *The challenge of diagnostic testing: Do we know what we are measuring?* Plenary presented at the annual meeting of the Language Testing Research Colloquium (LTRC), Ottawa, Canada.

Alderson, J. C. (2005b). *Diagnosing foreign language proficiency.* London: Continuum.

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK: Cambridge University Press.

Alderson, J. C. (1990a). Testing reading comprehension skills (Part One). *Reading in a Foreign Language, 6,* 425-438.

Alderson, J. C. (1990b). Test reading comprehension skills (Part Two). *Reading in a Foreign Language, 7,* 465-503.

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5,* 253-270.

Allan, A. (1992). *EFL reading comprehension test validation: Investigating aspects of process approaches.* Unpublished PhD thesis, Lancaster University.

Anderson, R. C. (1972). How to construct performance tests to assess comprehension. *Review of Educational Research, 42,* 145-170.

Anderson, R. C. (1982). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42,* 145-170.

Anderson, N. (1991). Individual differences in strategy use in second language reading and testing. *Modern Language Journal, 75,* 460-472.

Anderson, N., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data resources. *Language Testing, 8,* 41-66.

Anderson, R. C., & Pearson, P. D. (1988). A schema-theoretic view of basic processes in reading comprehension. In P. L. Carrell, J. Devine, & D. E. Eskey (eds.). *Interactive Approaches to Second Language Reading.* Cambridge: Cambridge University Press.

AnswerTree 3.1 User's Guide (2002). Chicago, IL: SPSS.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19,* 453-476.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13,* 125-150.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language.* Cambridge, UK: Cambridge University Press.

Bachman, L. F., Kunnan, A., Vanniarajan, S., & Lynch, B. (1988). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing, 5,* 128-59.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bernhardt, E. (2003). Challenges to reading research from a multilingual world. *Reading Research Quarterly, 38,* 112-117.

Bernhardt, E. (2000). Second-language reading as a case study of reading scholarship in the 20th Century. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 791-811). Mahwah, NJ: Lawrence Erlbaum.

Block, E. (1992). See how they read: Comprehension monitoring of L1 and L2 readers. *TESOL Quarterly, 26,* 319-341.

Block, E. (1986). The comprehension strategies of second language readers. *TESOL Quarterly, 20,* 463-494.

Bormuth, J. R. (1969). An operational definition of comprehension instruction. In K. S. Goodman & J. T. Fleming (Eds.), *Psycholinguistics and the teaching of reading.* Newark, Del: International Reading Association.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19,* 369-394.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47,* 423-466.

Butcher, K. R., & Kintsch, W. (2003). Text comprehension and discourse processing. In A. F. Healy & R. W. Proctor (Eds.), *Handbook of psychology: Experimental psychology* (pp. 575-595). New York: John Wiley & Sons, Inc.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1-47.

Carr, N. T. (2003). *An investigation into the structure of text characteristics and reader abilities in a test of second language reading.* Unpublished Ph. D. dissertation, University of California, Department of Applied Linguistics, Los Angeles, USA.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly, 19,* 727-752.

Carrell, P. L. (1984). The effects of rhetorical organization on ESL readers. *TESOL Quarterly, 17,* 441-469.

Carrell, P. L. (1983a). Some issues in studying the role of schemata, or background knowledge, in second language comprehension. *Reading in a Foreign Language, 1,* 81-92.

Carrell, P. L. (1983b). Three components of background knowledge in reading comprehension. *Language Learning, 33,* 183-203.

Carrell, P. L., & Grabe, W. (2002). Reading. In N. Schmitt (Ed.), *An Introduction to Applied Linguistics* (pp. 233-250). London: Arnold.

Carver, R. P. (1992a). Effect of prediction activities, prior knowledge, and text type upon amount comprehended: Using rauding theory to critique schema theory research. *Reading Research Quarterly, 27,* 165-174.

Carver, R. P. (1992b). What do standardized tests of reading comprehension measure in terms of efficiency, accuracy, and rate? *Reading Research Quarterly, 27,* 347-359.

Cheng, L. (2003). *Academic reading strategies used by Chinese EFL learners: Five case studies.* Unpublished PhD thesis, University of British Columbia.

Chomsky, N. (1968). *Language and mind.* New York, NY: Harcourt, Brace & World.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background knowledge on reading comprehension.* Cambridge, UK: University of Cambridge Local Examinations Syndicate.

Cohen, A. (1998). *Strategies in learning and using a second language.* London: Longman.

Cohen, A. (2005, July). *The coming of age of test-taking strategies.* Symposium presented at the annual conference of the Language Testing Research Colloquium, Ottawa, Canada.

Cohen, J., & Kolstad, A. (2000, April). *Theory-consistent item response models.* Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks.* TOEFL Monograph Report No. MS-33.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system.* [Computer software]. Iowa City, IA: ACT.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Experimental Education, 56,* 67-76.

DiBello, L. V., & Crone, C. (2001, April). *Technical methods underlying the PSAT/NMSQTTM enhanced score report.* Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment.* (pp. 361-389). Hillsdale, NJ: Erlbaum.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. R*eading research quarterly, 16,* 486-514.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150).

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38,* 343-368.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175-93.

English Language Institute. (2003). *Michigan English Language Assessment Battery Technical Manual.* MELAB Program, University of Michigan, Ann Arbor, MI.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied measurement in education, 15*, 49-74.

Enright, M. K., Grabe, W., Koda, D., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series MS-17). Princeton, NJ: Educational Testing Service.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Ewing, M., & Huff, K. (2004, April). *Using item difficulty modeling to evaluate skill categories.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Farhady, H., & Hessamy, G. (2005). *An empirical investigation of the L2 reading comprehension skills.* Paper presented at the annual meeting of the Language Testing Research Colloquium, Ottawa, CANADA.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 209-226.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 36,* 359-374.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty: Implications for construct validity. *Language Testing, 10,* 131-170.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing, 16,* 2-32.

Gantzer, J. (1996). Do reading tests match reading theory? *College ESL, 6,* 29-48.

Gao, L. (2002). *Passage dependence of the reading comprehension component of the College English Test.* Unpublished master's thesis, Queen's University, Kingston, Ontario, CANADA.

Gitomer, D. H., & Rock, D. (1993). Addressing process variables in test analysis. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 243-268). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Journal of the reading specialist, 6,* 126-135.

Gorin, J. S. (2002). *Cognitive and psychometric modeling of text-based reading comprehension GRE-V items.* Unpublished Ph. D. dissertation, University of Kansas, Department of Psychology, Kansas, USA.

Grabe, W. (2004). Research on teaching reading. *Annual Review of Applied Linguistics, 24,* 44-69.

Grabe, W. (2002). Reading in a second language. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 49-59). New York: Oxford University Press.

Grabe, W. (1999). Developments in reading research and their implications for computer-adaptive reading assessment. In M. Chalhoub-Deville (Ed.), *Issues in computer adaptive testing of reading proficiency* (pp. 11-48). Cambridge, UK: University of Cambridge Local Examinations Syndicate.

Grabe, W. (1991). Current development in second-language reading research. *TESOL Quarterly, 25,* 375-406.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading.* London, UK: Pearson Education.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff Publishing.

Hambleton, R, K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* London, UK: SAGE Publications.

Hill, C., & Parry, K. (1992). The test at the gate: Models of literacy in reading assessment. *TESOL Quarterly, 26,* 433-461.

Hudson, T. (1998). Theoretical perspectives on reading. *Annual Review of Applied Linguistics, 18,* 43-60.

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000* (TOEFL Monograph Series MS-4). Princeton, NJ: Educational Testing Service.

Huff, K. (2003). *An item modeling approach to providing descriptive score reports.* Unpublished doctoral disseration, University of Massachusetts Amherst.

Hymes, D. (1972). Models of interaction of language and social life. In J. J. Gumperz & D. Hymes (eds.), *Directions in sociolinguistics: The ethnography of communication* pp. 35-71. New York: Holt, Rinehart and Winston.

Introduction to AnswerTree (2002). Chicago, IL: SPSS.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TORFL 2000 framework: A working paper* (TOEFL Monograph Series MS-16). Princeton, NJ: Educational Testing Service.

Johnston, P. (1983). *Reading comprehension assessment: A cognitive basis.* Newark, Delaware: International Reading Association

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly, 21,* 220-39.

Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the Test of English as a Foreign Language (TOEFL).* Unpublished PhD dissertation, University of Illinois at Urbana-Champaign.

Katz, S., & Lautenschlager, G. J. (2001). The contribution of passage and no-passage factors to item performance on the SAT reading task. *Educational Assessment, 7,* 165-176.

Katz, S., & Lautenschlager, G. J. (1995). The SAT reading task in question: Reply to Freedle and Kostin. *Psychological Science, 6,* 126-127.

Keppel, G., & Zedeck, S. (2001). *Data analysis for research designs: analysis of variance and multiple regression/correlation approaches.* New York: W. H. Freeman.

Kerlinger, F.N. (1979). *Behavioral research: A conceptual approach.* New York: Holt.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge: Cambridge University Press.

Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25,* 5-30.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* New York, NY: Cambridge University Press.

Koda, K. (1996). L2 word recognition research: a critical review. *The Modern Language Journal, 80,* 450-460.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* New York, NY: Springer.

Kolln, M. (1999). *Rhetorical grammar: Grammatical choices, rhetorical effects* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293-323.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing, 21,* 74-100.

Leighton, J. P., & Gierl, M. (2005). *Identifying models of cognition in educational measurement.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, London, Ontario, Canada.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(4), 6-15.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of*

Educational Measurement, 41, 205-237.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lumley, T., & McNamara, T. F. (1995). Rater Characteristics and Rater Bias: Implications for Training. *Language Testing, 12*(1), 54-71.

Lumley, T., & Brown, A. (2004). Test-taker and rater perspectives on integrated reading and writing tasks in the Next Generation TOEFL. *Language Testing Update, 35,* 75-79.

Lunzer, E., & Gardner, K. (1979). *The effective use of reading (Eds.).* London: Heinemann Educational Books.

McKeown, M. G., Beck, I. L., Sinatra, G. M., & Losterman, J. A. (1992). The contribution of prior knowledge and coherent text to comprehension. *Reading Research Quarterly, 27,* 79-93.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on uncertainty in artificial intelligence* (pp. 437-446). San Francisco: Morgan Kaufmann.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software Inc.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19,* 477-496.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

Munby, J. (1978). *A communicative syllabus design: A sociolinguistic model for defining the content of purpose-specific language programmes.* New York, NY: Cambridge University Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64,* 575-603.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19,* 395-418.

Paris, S. G., Wasik, B. A., & Turner, J. C. (1991). The development of strategic readers. In R. Barr, M. Kamil, P. Mosenthal, & P. D. Pearson, (Eds.), *Handbook of reading* *research* (Vol. 1, pp. 609-640). White Plains, NY: Longman.

Pearson, P. D., & Johnston, D. D. (1978). *Teaching reading comprehension.* New York: Holt, Rinehart and Winston.

Perfetti, C. A. (1995). Cognitive research can inform reading education. *Journal of Research in Reading, 18,* 106-115.

Perkins, K., & Brutten, S. R. (1992). The effect of processing depth on ESL reading comprehension. *Journal of Research in Reading, 15,* 67-81.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20,* 26-56.

Phillips, L., M. (1988). Young readers' inference strategies in reading comprehension. *Cognition and Instruction, 5,* 193-222.

Phillips, L. M., & Norris, S. P. (2002). Schema theory criticisms. In B. J. Guzzetti (Ed.), *Literacy in America: An encyclopedia of history, theory, and practice* (pp. 558-561). Santa Barbara, CA: ABC-CLIO.

Powers, D. E., & Leung, S. W. (1995). Answering the new SAT reading comprehension questions without the passages. *Journal of Educational Measurement, 32,* 105-129.

Puhan, G. (2003). *Evaluating the effectiveness of two-stage testing for English and French examinees on the SAIP Science 1996 and 1999 tests.* Unpublished PhD dissertation, University of Alberta.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review, 56,* 282-307.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice Hall.

Radford, A. (2004). *English Syntax: An introduction.* New York, NY: Cambridge University Press.

Roller, C. (1990). Commentary: The interaction of knowledge and structure variables in the processing expository prose. *Reading Research Quarterly, 25,* 79-89.

Ruddell, R. B., & Ruddell, M. R., & Singer, H. (1994). *Theoretical models and processes of reading.* Newark, Delaware: International Reading Association.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Domic (Ed.), *Attention and performance* (pp. 28-59). UL, NY: Academic Press.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education* (pp. 33-58). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension tests. *International Journal of Testing, 1,* 185-216.

Samejima, F. (1997). Graded response model. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). Ann

Arbor, MI: Edwards Brothers.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26,* 113-125.

Sheehan, K. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34,* 333-352.

Sheehan, K., & Ginther, A. (2001). *What do multiple choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on a standardized test of reading comprehension skill.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Singer, M., & Kintsch, W. (2001). Text retrieval: A theoretical exploration. *Discourse Processes, 31,* 27-59.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17,* 38-62.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (eds.), *Cognition and Second Language Instruction* (pp. 183-205). New York, NY: Cambridge University Press.

Smith, F. (1971). *Understanding reading: A psycholinguistic analysis of reading and learning to read.* New York: Holt, Rinehart and Winston.

Smith, F. (2004). *Understanding reading.* Hillsdale, NJ: Lawrence Erlbaum Associates.

SPSS (2005). *SPSS Version 13.0* [Computer Software]. Chicago, IL: SPSS.

Stanovich, K. E. (1980). Towards an interactive compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16,* 32-71.

Stanovich, K. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: Guilford Press.

Sternberg, R. J., & Ben-Zeev, T. (2001). *Complex cognition: The psychology of human thought.* New York, NY: Oxford University Press.

Strong-Krause, D. (2001). *English as a second language speaking ability: A study in domain theory development.* Unpublished PhD thesis, Brigham Young University.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

Thompson, G. (2004). *Introducing functional grammar* (2nd ed.). New York, NY: Oxford University Press.

Thorndike, R. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychological Association, 8,* 323-332.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice.* New York, NY: Addison Wesley Longman.

Van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 1-28). New York: Springer.

Van Essen, T. (2001, April). *Developing and presenting enhanced skill descriptors for the PSAT/ NMSQT*^{TM}. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

VanderVeen, A. (2004). *Toward a construct of critical reading for the new SAT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Wainer, H. & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational & Psychological Measurement, 57,* 741-58.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement, 37,* 113-140.

Wang, X., Bradlow, E. T., & Wainer , H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26,* 109 – 128.

Wang, X., Bradlow, E. T., & Wainer , H. (2004). *User's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis.* Princeton, NJ: Educational Testing Service; Philadelphia, PA: National Board of Medical Examiners.

Weir, C. J. (1983). *Identifying the language needs of overseas students in tertiary education in the United Kingdom.* Unpublished PhD thesis, University of London, Institute of Education.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrica, 45,* 479-494.

Williams, E., & Moran, C. (1989). Reading in a foreign language at intermediate and advanced levels with particular reference to English. *Language Teaching, 22,* 217-228.

Wu, Y. (1998). What do tests of listening comprehension test? A retrospective study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*, 21-44.

Xi, X. (2003). *Investigating language performance on the Graph Description Task in a semi-direct Oral Test.* Unpublished PhD dissertation, University of California, Los Angeles.

Yang, P. (2000). *Effects of test-wiseness upon performance on the Test of English as a Foreign Language.* Unpublished PhD dissertation, University of Alberta, Edmonton, Alberta, Canada.

Appendix A: Recruiting Letter and Letter of Information for Raters

Recruiting Letter

Dear Graduate Students in Educational Psychology:

I am a doctoral student of the Department of Educational Psychology at the University of Alberta. I am conducting a study modeling the cognitive processes underlying performance on the reading items included in the Michigan English Language Assessment Battery (MELAB). The MEALB is a large-scale high stakes assessment of English as a foreign language and its assessment results are widely used by educational institutions, government agencies, and licensing agencies to make decisions about educational and employment opportunities.

This study is part of my studies for a doctoral degree and the information you provide will help identify the cognitive processes required to correctly answer the MELAB reading test items. Your participation would be greatly appreciated!

The data will be collected within a week in end November or early December 2005 and participating in this study will take a total of ten hours or so of your time. To participate in this study, you are expected to have expertise in the domain of L2 reading (e.g., took courses in applied linguistics or related area) and experience of teaching reading to adult EFL learners.

For details about participating this study, please refer to the letter of information attached in this email. Thank you for your consideration!

Sincerely,
Lingyun Gao, PhD candidate

Letter of Information

Dear Graduate Students in Educational Psychology:

I am a doctoral student of the Department of Educational Psychology at the University of Alberta. I am conducting a study modeling the cognitive processes underlying performance on the reading items included in the Michigan English Language Assessment Battery (MELAB), a large-scale high stakes assessment of English as a foreign language. This study is part of my studies for a doctoral degree and the information you provide will help identify the cognitive processes required to correctly answer the MELAB reading test items. Your participation would be greatly appreciated!

The data collection procedure consists of four steps, which will take a total of ten hours or so of your time. First, you will participate in a one-hour group training session to be held by the researcher at the University of Alberta to introduce you to the MELAB test forms, rating instrument, and rating procedures. Second, you will use a rating instrument to independently code the 40 items included in two forms of the MELAB reading test in terms of the cognitive processes required to correctly answer each item. You will be given three days to complete the task and return your completed work with all materials to the researcher by 5pm of the third day. Third, after coding, you will participate in a meeting, during which you will look at the summary of the coding conducted by all the raters and reach consensus where disagreement occurs. Fourth, after the verbal report data are collected from the students and coded in phase 2 of the study, you and I will meet as a group to compare the rating results and the processes actually used by the students to determine the cognitive processes required by the MELAB reading items. There are no risks or harms to you from participating in this study.

No names will be used in the data or published work; only codes (e.g., JR, MM) will be used in the data to identify the participants and only group results will be used. The data will not be released at any time to anyone, and will be used for research purposes only. The data will be sealed in envelopes and locked in the researcher's office. Only the researcher can access the data. Five years after completing the study, all the data regarding this study will be destroyed. The results of the study will be reported in the working paper to be submitted to the University of Michigan and will also be reported in the dissertation currently undertaken by the researcher. Results may also be reported in scholarly conferences and journals. The data for all uses will be handled in compliance with the University of Alberta Standards for the Protection of Human Research Participants. If you wish to have a copy of the report, you can contact me and I would be pleased to send you a copy. Your participation is voluntary. If you agree to participate, please be advised that due to the nature of this study (i.e., your rating results are vital for the later phases of the study), you may not be allowed to withdraw once the formal rating procedure begins. I would appreciate your foreseeing the possible events that may lead to your withdrawal from the study and advising me before that point so that I could find another rater.

Should you have any questions or concerns about this study, you may contact Lingyun Gao, the researcher, at gaog@ualberta.ca or 4925427, or Dr. Todd Rogers, supervisor of the researcher, at todd.rogers@ualberta.ca or 4923763. If you agree to participate in the study, please email me and I will contact you to set up the appointments. Thank you for your consideration!

Sincerely,
Lingyun Gao, PhD candidate

The plan for this study has been reviewed for its adherence to ethical guidelines and approved by the Faculties of Education, Extension and Augustana Research Ethics Board (EEA REB) at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Chair of the EEA REB at (780) 492-3751.

Appendix B: Rating Instrument

Directions:

This rating instrument contains four parts: Part I (The Rating Guide), Part II (The Rating Forms), Part III (Evidence of Ratings), and Part IV (Comments). Table 1 provides a list of cognitive processes that may be related to correctly answering the multiple-choice items included in the MELAB reading tests. To clarify the concepts involved in the rating process, Table 1 also provides definitions of the cognitive processes covered, rationale, and examples for your reference. Please code each item using the variables listed in the last column of Table 1, record your coding on the rating forms, and provide evidence for your ratings by taking notes, listing the key words, or bracketing the text containing the relevant information during rating. If you think correctly answering an item involves the cognitive processes not included in Table 1, please specify them on the rating forms under the column "Other".

Part I: The Rating Guide (Table 1 in Chapter 3 was attached in the next pages).

Part II: Rating Forms[1] (See the pages following Table 1)

Part III: Evidence for your ratings for the variables "Word Recognition Required" and "Vocabulary Knowledge Required" (See attached)

Part IV: Your comments on additional cognitive processes involved (See attached)

---

[1] As the Rating scales for Test Form E and Test Form F are exactly the same, only the Rating scales for Test Form E are included in this appendix. The Rating scales for Test Form F are identical with those for Test Form E.

Rating Form[1] (Test Form E)

Rater Code: ____________________ **Date** ____________________

Item	R1[2]	R2[3]	R3	R4	R5	R6	R7	T1	T2	T3	T4	TW1	TW2	TW3	Other[4]
1															
2															
3															
4															
5															
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															

Note. [1] Abbreviations used in the table: R1 = Word Recognition Required; R2 = Vocabulary Knowledge Required; R3 = Syntactic Knowledge Required; R4 = Knowledge of Discourse Structure Required; R5 = Synthesis Required; R6 = Inference Required; R7 = Purpose of Information; T1 = Location of Information; T2 = Type of Match; T3 = Number of Plausible Distractors; T4 = Topical Knowledge Required; TW1 = Item Cues; TW2 = Guess; TW3 = Surface Feature of Options.

[2] In the part of the text containing the information requested to correctly answer each item, please list any word(s) which need to be recognized using advanced phonological and/or orthographical knowledge, or the meaning of which need to be understood using context cues.

Item 1 ____________________

Item 2 ____________________

Item 3 ____________________

Item 4 ____________________

Item 5 ____________________

Item 6 ____________________

Item 7 ____________________

Item 8 ____________________

Item 9 ____________________

Item 10 ____________________

Item 11 ____________________

Item 12 ____________________

Item 13 ____________________

Item 14 ____________________

Item 15 ____________________

Item 16 ____________________

Item 17 ____________________

Item 18 ____________________

Item 19 ____________________

Item 20 ____________________

[3]Please list the infrequently used and specialized words in the part of the text containing the information requested to correctly answer each item.

Item 1 ______________________________

Item 2 ______________________________

Item 3 ______________________________

Item 4 ______________________________

Item 5 ______________________________

Item 6 ______________________________

Item 7 ______________________________

Item 8 ______________________________

Item 9 ______________________________

Item 10 ______________________________

Item 11 ______________________________

Item 12 ______________________________

Item 13 ______________________________

Item 14 ______________________________

Item 15 ______________________________

Item 16 ______________________________

Item 17 ______________________________

Item 18 ______________________________

Item 19 ______________________________

Item 20 ______________________________

[4]Please use the space below to describe other cognitive processes involved in obtaining the correct answer to each item.

Item 1 ____________________

Item 2 ____________________

Item 3 ____________________

Item 4 ____________________

Item 5 ____________________

Item 6 ____________________

Item 7 ____________________

Item 8 ____________________

Item 9 ____________________

Item 10 ____________________

Item 11 ____________________

Item 12 ____________________

Item 13 ____________________

Item 14 ____________________

Item 15 ____________________

Item 16 ____________________

Item 17 ____________________

Item 18 ____________________

Item 19 ____________________

Item 20 ____________________

Appendix C: Consent Form for the Raters

Project Title: Cognitive-Psychometric Modeling of the MELAB Reading Items

I, ___________________________, agree to participate in the study being conducted by the researcher. I have read the letter of information. Any questions that I had were answered to my satisfaction. I understand that I will be participating in a study modeling the cognitive processes underlying performance on the items included in the MELAB reading tests. I have been informed that my involvement consists of (1) participating in a one-hour group training session, (2) independently coding the 40 items included in two forms of the MELAB reading tests in the subsequent three days, and returning the coding results with all materials to the researcher by 5pm of the third day, (3) participating in a meeting to be scheduled on the following day to look at the summary of the coding conducted by all the raters and reach consensus where disagreement occurs, and (4) comparing the rating results to the processes used by the students to determine the features of the MELAB reading items. I understand that participating in this study will take a total of ten hours or so of my time and the information I provide will help to identify the cognitive processes required to correctly answer the MELAB reading test items. I understand that my participation is voluntary, and that due to the nature of this study, I am not allowed to withdraw once the formal rating procedure begins. I have been assured that my name will not appear on the data or published work to maintain my anonymity and confidentiality and that only codes (e.g., JR, MM) will be used in the data to identify the participants and only group results will be used. I am aware that I can contact the researcher and the researcher's supervisor at the University of Alberta with any questions, concerns, or complaints that I have.

By signing below, I certify that I have read the Consent Form.

Name of the Participant: ____
(Please print)

Signature of the Participant: ____

Date: ____

The plan for this study has been reviewed for its adherence to ethical guidelines and approved by the Faculties of Education, Extension and Augustana Research Ethics Board (EEA REB) at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Chair of the EEA REB at (780) 492-3751.

Appendix D: Procedures for Rater Training and MELAB Reading Item Coding

1. Training of the three raters

— The researcher provides 1-hour group training on day 1. The procedures are:

- Introduce the study and the MELAB reading test forms
- Acquaint the raters with the rating instrument:
    - ✓ Give them 10-15 minutes to read and review Table 1
    - ✓ The researcher describes each row of the table
    - ✓ The researcher describes each column of the table
    - ✓ Review together the scored variables in the last column of the table
    - ✓ Describe the rating form
- Clarify the rating procedure
- Encourage discussions to achieve common definitions & understanding of the procedure
- Provide the sample passage with five accompanying items for practice:
    - ✓ The raters first answer the set of sample items
    - ✓ The researcher provides the answer key for the raters to mark their answers
    - ✓ The raters code the items in term of all possible cognitive processes required to correctly answer each item, using the last column of Table 1 and the rating form.
    - ✓ Upon completion, the coding results are discussed, the rationale for coding particular cognitive processes shared, and inconsistencies resolved.

2. Formal data collection

— The raters independently code the MELAB reading items on days 2, 3, and 4.

General instructions for the rating procedures

Instructions: Each of you will be provided three envelopes, which contain the instructions and materials for each step of the rating task. Specifically,

Envelope A contains Form E and Form F of the MELAB reading test
Envelope B contains the answer keys to the 20 items included in each test form
Envelope C contains Table 1, rating form, and instructions for coding each item in terms of the cognitive processes required to correctly answer the corresponding item.

Please sequentially complete the task specified in each envelope. Do not proceed to open the next envelope until you complete the tasks in the previous envelope. You will be given three days to complete the entire task. Once you finish the tasks, please seal your completed work with all instructions and materials in the original envelopes and return them to me by 5pm of the third day. Thank you!

Instructions for the task specified in each envelope

Envelope A —

Instructions:

Please read through the two forms of the test paper and answer each item to the best of your ability, as if you were indeed taking a reading test.

Do not open envelope B until you complete answering all the items included in both forms.

Envelope B —

Instructions:

Now check your answers using the answer keys provided in this envelope. Score your test and correct your answers as necessary.

Upon completion, please proceed to open Envelope C.

Envelope C —

Instructions:

Table 1 in this envelope provides a list of cognitive processes that may be related to correctly answering the multiple-choice items included in the MELAB reading tests. To clarify the concepts involved in the rating process, Table 1 also provides definitions of the cognitive processes covered, rationale, and example cognitive-based item features for your reference.

Please open Envelope A and take out the two forms that you have completed. Now, code each item included in the two forms using the variables listed in the last column of Table 1 and record your coding on the rating form provided in this envelope. If you think correctly answering an item involves other cognitive processes, please specify them on the rating form under the column "Other". Should you have further questions, please contact me.

3. Coding summary

The researcher summarizes the item coding submitted by the three raters on the evening of day 4 (i.e., Thursday evening).

4. Group meeting

The researcher and the three raters will meet as a group on day 5 (Friday) to look at the item coding summarized by the researcher. Consistencies will be checked. Inconsistencies will be discussed to reach consensus item coding.

Appendix E: Summary of Item Ratings by the Three Raters[1]

Process		R1[2]			R2			R3			R4			R5			R6			R7		
Rater		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Item	E1	1	2	2	1	1	2	2	2	2	1	2	1	1	2	1	2	2	2	1	1	1
	E2	2	2	2	1	1	1	2	2	2	0	0	0	0	1	0	1	1	0	1	1	1
	E3	2	2	0	1	1	2	0	1	1	2	2	2	2	2	2	2	2	2	5	5	5
	E4	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	4	4	4
	E5	0	2	1	2	2	2	2	2	2	0	2	0	0	1	0	2	2	2	4	4	4
	E6	2	2	1	2	2	1	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1
	E7	1	2	1	1	1	0	2	1	1	0	0	0	0	0	0	1	1	1	1	1	1
	E8	2	2	1	0	1	1	0	1	2	1	2	1	1	2	2	2	2	1	4	2	2
	E9	1	0	0	1	1	1	0	1	0	0	1	0	1	0	1	0	0	0	1	1	1
	E10	0	0	0	1	1	1	0	0	2	0	0	0	0	0	2	2	2	2	5	5	4
	E11	0	2	2	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	4	4
	E12	0	1	1	2	2	1	0	2	2	0	0	1	0	0	0	0	1	1	1	2	2
	E13	1	1	1	2	1	2	1	1	1	2	2	0	0	1	1	0	1	0	4	4	4
	E14	2	2	2	2	2	2	0	2	2	0	2	2	0	1	1	1	1	1	3	3	2
	E15	1	1	2	2	2	1	1	2	1	0	0	0	0	1	0	1	1	1	5	5	4
	E16	2	2	2	2	2	2	2	2	2	0	1	0	0	1	1	1	1	0	4	4	4
	E17	1	2	1	0	1	1	0	1	2	1	0	2	1	1	2	1	1	1	4	4	2
	E18	0	2	1	1	0	0	2	2	2	1	1	1	0	1	0	1	1	1	4	4	4
	E19	1	1	2	0	0	1	2	2	2	1	2	1	1	2	1	1	2	2	1	1	4
	E20	0	1	1	0	0	0	2	2	2	0	1	0	0	1	1	1	1	0	4	4	4

Appendix E (*continued*)

Process		T1			T2			T3			T4			TW1			TW2			TW3		
Rater		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Item	E1	2	2	2	1	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	E2	3	3	3	2	2	2	3	3	1	0	0	0	0	0	0	0	0	0	0	0	0
	E3	4	4	4	3	3	3	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0
	E4	2	2	2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	E5	1	1	1	1	1	1	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	E6	2	2	2	2	2	2	3	1	2	0	0	0	0	0	0	0	0	0	0	0	0
	E7	2	2	2	1	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	E8	3	4	1	3	3	3	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	E9	3	3	2	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	E10	4	3	4	2	2	3	2	3	2	1	0	0	0	0	0	0	0	0	0	0	0
	E11	3	3	3	2	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	E12	2	2	1	2	3	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	E13	2	2	2	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	E14	1	1	1	1	1	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	E15	2	3	2	2	2	2	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	E16	3	3	3	2	2	3	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	E17	3	3	3	2	2	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0
	E18	3	2	3	2	3	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	E19	2	2	2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	E20	1	1	1	1	1	1	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0

Appendix E (*continued*)

Process		R1^2			R2			R3			R4			R5			R6			R7		
Rater		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Item	F1	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	0	0	0	1	1	4
	F2	1	2	2	1	2	1	2	2	2	0	1	1	1	1	1	1	1	1	1	1	4
	F3	0	1	1	1	1	0	2	2	2	0	0	0	0	0	0	0	0	0	1	1	1
	F4	1	1	1	2	1	1	1	1	1	0	0	0	0	0	0	1	1	1	4	4	4
	F5	0	1	1	1	1	0	1	1	1	0	0	0	0	1	1	0	0	0	4	3	3
	F6	2	1	1	0	0	1	1	1	1	2	2	2	2	2	2	2	2	2	1	1	4
	F7	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	2	1	2	4	4	4
	F8	2	2	2	1	1	1	2	2	2	1	2	2	1	1	1	1	1	1	5	5	5
	F9	1	2	1	0	1	0	1	2	1	2	2	2	2	2	1	1	0	1	3	5	5
	F10	1	2	1	2	1	2	1	2	1	1	1	0	1	1	0	1	1	0	4	4	4
	F11	2	1	2	1	1	1	1	1	2	2	2	2	2	2	2	1	2	2	3	3	3
	F12	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	2	2	1	4	4	4
	F13	2	2	2	2	1	2	1	2	2	2	1	1	2	2	2	2	2	2	4	4	1
	F14	2	2	2	1	2	0	1	2	2	0	2	2	2	2	0	0	2	1	4	4	4
	F15	2	2	2	1	1	2	1	2	2	0	1	1	2	2	1	2	1	1	4	4	4
	F16	1	2	1	0	1	1	1	1	2	1	1	1	1	2	2	1	1	2	3	3	3
	F17	0	1	0	0	0	0	2	2	2	0	0	0	1	0	0	0	0	0	4	4	4
	F18	2	2	2	2	2	1	2	2	2	0	1	1	1	1	1	2	2	1	2	2	2
	F19	2	2	2	0	1	1	1	1	2	2	2	1	2	2	1	0	2	2	2	2	3
	F20	2	2	1	2	0	1	2	1	2	1	0	0	1	1	1	1	1	2	1	1	4

Appendix E (*continued*)

Process		T1			T2			T3			T4			TW1			TW2			TW3		
Rater		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Item	F1	2	4	2	3	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	F2	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	F3	2	2	2	2	3	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	F4	3	3	3	3	3	3	3	3	2	0	0	0	0	0	0	0	0	0	0	0	0
	F5	3	3	3	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	F6	3	4	4	3	3	3	2	3	3	0	0	0	0	0	0	0	0	0	0	0	0
	F7	3	4	4	2	2	2	3	3	1	0	0	0	0	0	0	0	0	0	0	0	0
	F8	2	2	2	2	2	2	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
	F9	1	1	1	2	2	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	F10	3	3	3	2	2	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	F11	4	4	4	3	3	3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	F12	1	2	1	2	2	2	3	3	2	0	0	0	0	0	0	0	0	0	0	0	0
	F13	4	3	3	3	2	2	3	2	2	0	0	0	0	0	0	0	0	0	0	0	0
	F14	2	2	2	1	1	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	F15	3	3	3	3	2	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	F16	3	3	3	3	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	F17	3	3	3	2	2	2	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0
	F18	2	2	2	2	2	2	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0
	F19	1	1	1	3	2	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	F20	3	3	3	1	3	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0

Note. [1]The first row contains 14 cognitive variables. The second row contains the three raters (i.e., Rater A, Rater B, and Rater C). The next rows contain the ratings by each rater for the items included in Forms E and F. As seen in the table, each of the three raters coded a total of 40 items on a list of 14 variables.

[2]Abbreviations used in the table: R1 = Word Recognition Required; R2 = Vocabulary Knowledge Required; R3 = Syntactic Knowledge Required; R4 = Knowledge of Discourse Structure Required; R5 = Synthesis Required; R6 = Inference Required; R7 = Purpose of Information; T1 = Location of Information; T2 = Type of Match; T3 = Number of Plausible Distractors; T4 = Topical Knowledge Required; TW1=Item Cues; TW2 = Guess; TW3 = Surface Feature of Options.

Appendix F: Recruiting Letter and Letter of Information for Verbal Participants

Recruiting Letter

Dear Students:

I am a doctoral student of the Department of Educational Psychology at the University of Alberta. I am conducting a study modeling the cognitive processes underlying performance on the reading items included in the Michigan English Language Assessment Battery (MELAB). The MEALB is a large-scale high stakes assessment of English as a foreign language and its assessment results are widely used by educational institutions, government agencies, and licensing agencies to make decisions about educational and employment opportunities.

This study is part of my studies for a doctoral degree and the information you provide will help identify the cognitive processes used to correctly answer the MELAB reading test items. Your participation would be greatly appreciated!

The data will be collected during the period from Fall 2005 to Winter 2006 and participating in this study will take approximately 2.5 hours of your time. To participate in the study, you are expected to be Chinese-speaking students who started an undergraduate or graduate program at the University of Alberta in fall 2005 and have resided in Canada for no longer than nine months.

For further details, please refer to the letter of information attached in this email. If you agree to participate, please simply reply to this email and I will contact you to set up the appointment. Thank you very much for your consideration!

Sincerely,
Lingyun Gao, PhD Candidate

Letter of Information

Dear Students:

I am a doctoral student of the Department of Educational Psychology at the University of Alberta. I am conducting a study modeling the cognitive processes underlying performance on the reading items included in the Michigan English Language Assessment Battery (MELAB), a large-scale high stakes assessment of English as a foreign language. This study is part of my studies for a doctoral degree and the information you provide will help identify the cognitive processes used to correctly answer the MELAB reading test items. Your voluntary participation would be greatly appreciated.

The data collection procedure consists of two separate sessions. On day 1, I will ask you

Appendix G: Procedures and Instructions for Verbal Report Sessions

[Setting: The researcher and the participant sit side-by-side at a table on which there is a digital audio recorder, a microphone, and a folder containing a consent form, a sheet of directions, a couple of worm-up tasks, and Form E or F of the MELAB reading tests.]

Session 1

I. The researcher explains the nature and procedure of the task:
Instruction:
Hello, my name is Lingyun Gao. I am a doctoral student in the Department of Educational Psychology at the University of Alberta. I am currently conducting a study modeling the cognitive processes underlying the MELAB reading test items. In this study, I am interested in the cognitive processes you use to answer the test items. You will be asked to verbally report your thought processes while answering the items and your remembrances about your thoughts after completing each item. You will also be asked to answer several questions of basic background information before we begin and to rate the topic familiarity for each passage after your verbal reports. Data collection will be conducted during two separate sessions within this week. We will complete a total of 10 items based on two passages during each session. Since I will be asking you to talk quite a bit during each session, I will be using a digital audio recorder to make sure that I capture everything that you tell me. This is completely voluntary and I want to be sure you are comfortable with being a part of this interview. Do you agree to participate? [If yes], Would you please fill out the consent form?
[The researcher provides the participant with the consent form to read and sign].

II. After the participant has read and signed the consent form, the researcher interviews the participant using the questions below:

1. *How old are you?*
2. *What is your level / years of education?*
3. *Could I know your discipline of study?*
4. *How long did you study English in your first country? (years and months)*
5. *How much time have you spent in an English speaking country?*

III. The researcher introduces the verbal reporting procedures and presents the warm-up tasks (see Appendix K) to familiarize the participant with the verbal reporting procedure.

IV. Data collection using the Form E or Form F of the MELAB reading test

Instruction:
In this main experiment, I'm going to ask you to read two passages and answer 10 multiple-choice items based on these passages. It is important that you read the passages and answer the questions as if you were taking a real reading test. As you work through the items, I would like you to think aloud as you did in the practice. That is, I want you to talk out aloud about everything that you are thinking and attending to in whatever language you are thinking, from the time you start reading the question stem until you select an answer. Please do not try to plan out or explain to me what you are thinking. It may help to imagine that you are in the room by yourself. It is important that you talk constantly. If you are silent for any long period of time, I will remind you to keep talking.

Once you have answered an item, I would like you to tell me alll that you can remember about what you were thinking and where you were attending to, from the time you began to read the question until you decide your answer, just as you did in the practice. I am interested in what you can actually remember, not what you think you may or should have thought. If possible, it would be best if you can tell me what you remember in the order in which your memories occurred as you worked through each item. If you are not sure about any of your memories, please say so. I do not want you to try to answer the question again. I just want you to tell me what you can remember about your thinking and the place you were attending to when you read and answered the item. You will not be interrupted or assisted once you begin, except that if you pause for any long period of time, you will be reminded to keep talking. Do you understand what I want you to do?

[If the participant has no more questions, the researcher administers the first two passages with their accompanying 10 items from Form E or F and turns on the digital recorder.]
Now, let's start with some reading test items.

[After they have completed the items and verbal reports]
Now, I would like you to rate your familiarity with the passage topics using the scale 0 = familiar, 1 = somewhat familiar, 2 = not familiar. The topic familiarity ratings are intended to determine whether topical knowledge is a source of processing difficulty that affects the difficulty of the MELAB reading items.

Session 2

Following the verbal report procedures described in IV, the participants complete the remaining two passages with their accompanying 10 items on Form E or Form F of the MELAB reading test.

Appendix H: Practice Tasks for the Verbal Reporting Sessions[1]

Instruction and Tasks for Practicing Concurrent and Retrospective Verbal Report Skills

In this experiment, I am interested in what you think about when you find answers to the questions that I am going to ask you to answer. I will ask you to *think aloud* as you work on the problem. That is, I want you to tell me *everything* you are thinking from the time you first see the question until you give an answer. I would like you to talk aloud *constantly* from the time I present each problem until you have given your final answer to the question. I don't want you to plan out what to say or explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. You can talk aloud in English, Chinese or both. It is most important that you keep talking. If you are silent for longer than 10 seconds, I will remind you to keep talking. Do you understand what I want you to do?

Good, now let's start with some practice problems. First, I want you to multiply two numbers in your head and tell me what you are thinking as you get an answer.

What is the result of multiplying 22 x 28?

Good! Now I want to see how much you can remember about what you were thinking from the time you read the question until you gave the answer. I am interested in what you actually can *remember* rather than what you think you must have thought. If possible, I would like you to tell me about your memories in the sequence in which they occurred while working on the question. Please tell me if you are uncertain about any of your memories. I don't want you to work on solving the problem again, just report all that you can remember thinking about when answering the question. Now tell me what you remember. You can tell me in English, Chinese, or both.

Good. Now I will give you two more practice problems before we proceed with the main experiment. I want you to do the same thing for each of these problems. I want you to think aloud as before while you think about the question, and after you have answered it, I will ask you to report all that you can remember about your thinking. Any questions? Here is your next problem.

How many windows are there in your parent's house?

Now tell me all that you can remember about your thinking.
Good, now here is another practice problem. Please think aloud as you try to answer it. There is no need to keep count, I will keep track for you.

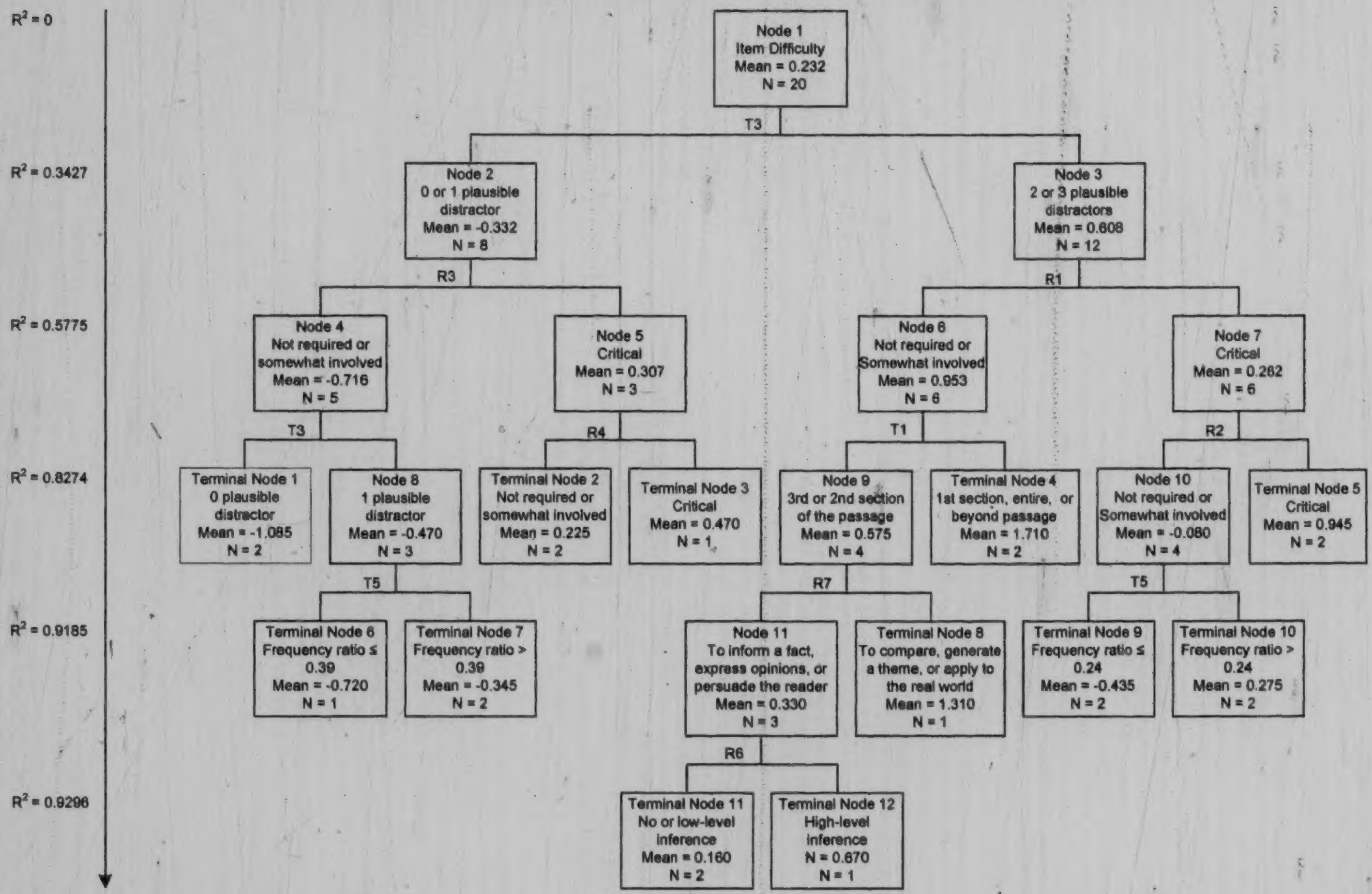
Name 20 animals.

Now tell me all that you can remember about your thinking.

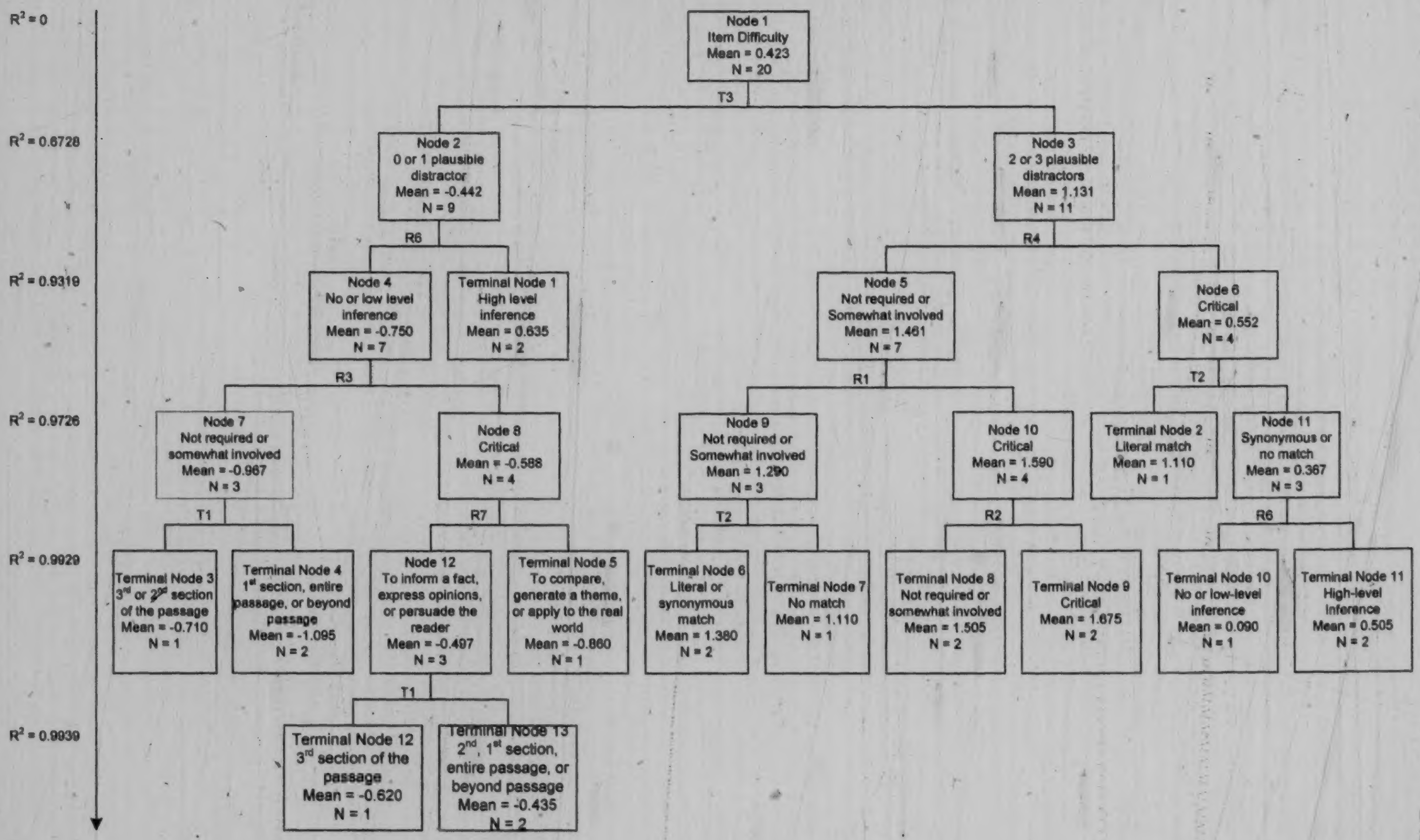
---

[1] Selected from Ericsson and Simon (1993).

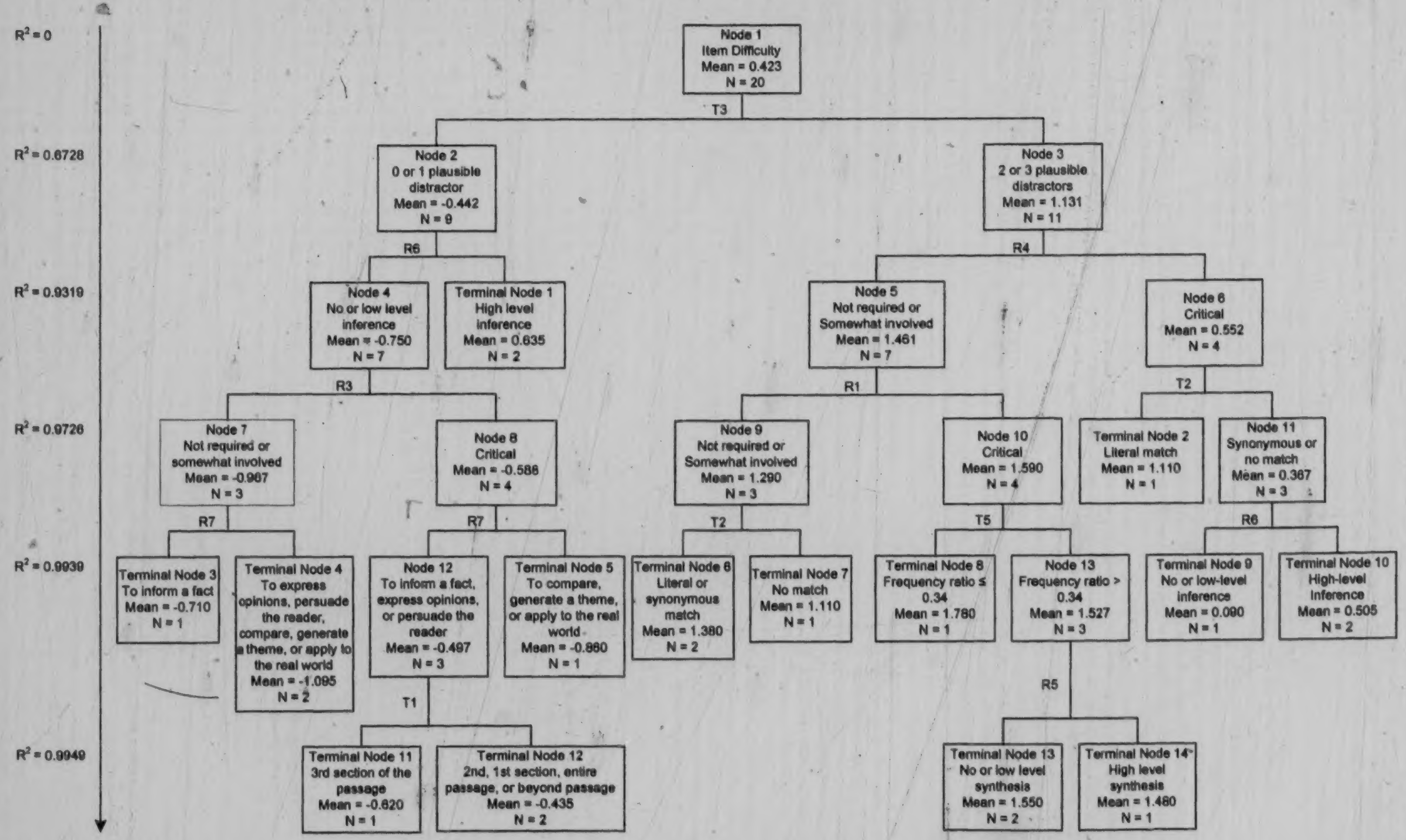
Appendix K: The Initial Tree for the 11-Component Cognitive Model (Form E)


$R^2 = 0$
$R^2 = 0.3427$
$R^2 = 0.5775$
$R^2 = 0.8274$
$R^2 = 0.9185$
$R^2 = 0.9296$
Node 1
Item Difficulty
Mean = 0.232
N = 20
T3
Node 2
0 or 1 plausible distractor
Mean = -0.332
N = 8
Node 3
2 or 3 plausible distractors
Mean = 0.608
N = 12
R3
R1
Node 4
Not required or somewhat involved
Mean = -0.716
N = 5
Node 5
Critical
Mean = 0.307
N = 3
Node 6
Not required or Somewhat involved
Mean = 0.953
N = 6
Node 7
Critical
Mean = 0.262
N = 6
T3
R4
T1
R2
Terminal Node 1
0 plausible distractor
Mean = -1.085
N = 2
Node 8
1 plausible distractor
Mean = -0.470
N = 3
Terminal Node 2
Not required or somewhat involved
Mean = 0.225
N = 2
Terminal Node 3
Critical
Mean = 0.470
N = 1
Node 9
3rd or 2nd section of the passage
Mean = 0.575
N = 4
Terminal Node 4
1st section, entire, or beyond passage
Mean = 1.710
N = 2
Node 10
Not required or Somewhat involved
Mean = -0.080
N = 4
Terminal Node 5
Critical
Mean = 0.945
N = 2
T5
R7
T5
Terminal Node 6
Frequency ratio ≤ 0.39
Mean = -0.720
N = 1
Terminal Node 7
Frequency ratio > 0.39
Mean = -0.345
N = 2
Node 11
To inform a fact, express opinions, or persuade the reader
Mean = 0.330
N = 3
Terminal Node 8
To compare, generate a theme, or apply to the real world
Mean = 1.310
N = 1
Terminal Node 9
Frequency ratio ≤ 0.24
Mean = -0.435
N = 2
Terminal Node 10
Frequency ratio > 0.24
Mean = 0.275
N = 2
R6
Terminal Node 11
No or low-level inference
Mean = 0.160
N = 2
Terminal Node 12
High-level inference
N = 0.670
N = 1

Appendix L: The Initial Tree for the 10-Component Cognitive Model (Form F)


R² = 0
R² = 0.6728
R² = 0.9319
R² = 0.9726
R² = 0.9929
R² = 0.9939
Node 1 Item Difficulty Mean = 0.423 N = 20
T3
Node 2 0 or 1 plausible distractor Mean = -0.442 N = 9
Node 3 2 or 3 plausible distractors Mean = 1.131 N = 11
R6
R4
Node 4 No or low level inference Mean = -0.750 N = 7
Terminal Node 1 High level inference Mean = 0.635 N = 2
Node 5 Not required or Somewhat involved Mean = 1.461 N = 7
Node 6 Critical Mean = 0.552 N = 4
R3
R1
T2
Node 7 Not required or somewhat involved Mean = -0.967 N = 3
Node 8 Critical Mean = -0.588 N = 4
Node 9 Not required or Somewhat involved Mean = 1.290 N = 3
Node 10 Critical Mean = 1.590 N = 4
Terminal Node 2 Literal match Mean = 1.110 N = 1
Node 11 Synonymous or no match Mean = 0.367 N = 3
T1
R7
T2
R2
R6
Terminal Node 3 3rd or 2nd section of the passage Mean = -0.710 N = 1
Terminal Node 4 1st section, entire passage, or beyond passage Mean = -1.095 N = 2
Node 12 To inform a fact, express opinions, or persuade the reader Mean = -0.497 N = 3
Terminal Node 5 To compare, generate a theme, or apply to the real world Mean = -0.860 N = 1
Terminal Node 6 Literal or synonymous match Mean = 1.380 N = 2
Terminal Node 7 No match Mean = 1.110 N = 1
Terminal Node 8 Not required or somewhat involved Mean = 1.505 N = 2
Terminal Node 9 Critical Mean = 1.675 N = 2
Terminal Node 10 No or low-level inference Mean = 0.090 N = 1
Terminal Node 11 High-level Inference Mean = 0.505 N = 2
T1
Terminal Node 12 3rd section of the passage Mean = -0.620 N = 1
Terminal Node 13 2nd, 1st section, entire passage, or beyond passage Mean = -0.435 N = 2

Appendix M: The Initial Tree for the 11-Component Cognitive Model (Form F)


R² = 0
R² = 0.6728
R² = 0.9319
R² = 0.9726
R² = 0.9939
R² = 0.9949
Node 1 Item Difficulty Mean = 0.423 N = 20
T3
Node 2 0 or 1 plausible distractor Mean = -0.442 N = 9
Node 3 2 or 3 plausible distractors Mean = 1.131 N = 11
R6
R4
Node 4 No or low level inference Mean = -0.750 N = 7
Terminal Node 1 High level inference Mean = 0.635 N = 2
Node 5 Not required or Somewhat involved Mean = 1.461 N = 7
Node 6 Critical Mean = 0.552 N = 4
R3
R1
T2
Node 7 Not required or somewhat involved Mean = -0.967 N = 3
Node 8 Critical Mean = -0.586 N = 4
Node 9 Not required or Somewhat involved Mean = 1.290 N = 3
Node 10 Critical Mean = 1.590 N = 4
Terminal Node 2 Literal match Mean = 1.110 N = 1
Node 11 Synonymous or no match Mean = 0.367 N = 3
R7
R7
T2
T5
R6
Terminal Node 3 To inform a fact Mean = -0.710 N = 1
Terminal Node 4 To express opinions, persuade the reader, compare, generate a theme, or apply to the real world Mean = -1.095 N = 2
Node 12 To inform a fact, express opinions, or persuade the reader Mean = -0.497 N = 3
Terminal Node 5 To compare, generate a theme, or apply to the real world Mean = -0.860 N = 1
Terminal Node 6 Literal or synonymous match Mean = 1.380 N = 2
Terminal Node 7 No match Mean = 1.110 N = 1
Terminal Node 8 Frequency ratio ≤ 0.34 Mean = 1.780 N = 1
Node 13 Frequency ratio > 0.34 Mean = 1.527 N = 3
Terminal Node 9 No or low-level inference Mean = 0.090 N = 1
Terminal Node 10 High-level inference Mean = 0.505 N = 2
T1
R5
Terminal Node 11 3rd section of the passage Mean = -0.820 N = 1
Terminal Node 12 2nd, 1st section, entire passage, or beyond passage Mean = -0.435 N = 2
Terminal Node 13 No or low level synthesis Mean = 1.550 N = 2
Terminal Node 14 High level synthesis Mean = 1.480 N = 1

Appendix N: Correlation Matrix (Form E)

		R1	R2	R3	R4	R5	R6	R7	T1	T2	T3	T5
R1	Pearson Correlation	1	.332	.280	.186	.253	.225	-.138	-.031	.225	.350	-.078
	Sig. (2-tailed)		.153	.233	.433	.281	.340	.560	.898	.340	.130	.742
	N	20	20	20	20	20	20	20	20	20	20	20
R2	Pearson Correlation	.332	1	.100	.005	-.092	-.170	.202	-.100	.006	-.067	.408
	Sig. (2-tailed)	.153		.675	.984	.699	.475	.394	.674	.980	.778	.074
	N	20	20	20	20	20	20	20	20	20	20	20
R3	Pearson Correlation	.280	.100	1	.213	-.021	.205	.045	-.511*	-.163	.293	.210
	Sig. (2-tailed)	.233	.675		.367	.931	.385	.849	.021	.492	.210	.375
	N	20	20	20	20	20	20	20	20	20	20	20
R4	Pearson Correlation	.186	.005	.213	1	.685**	.198	.083	.055	.053	-.183	-.215
	Sig. (2-tailed)	.433	.984	.367		.001	.403	.727	.818	.824	.439	.362
	N	20	20	20	20	20	20	20	20	20	20	20
R5	Pearson Correlation	.253	-.092	-.021	.685**	1	.254	.000	.345	.254	-.061	-.360
	Sig. (2-tailed)	.281	.699	.931	.001		.281	1.000	.136	.281	.798	.119
	N	20	20	20	20	20	20	20	20	20	20	20
R6	Pearson Correlation	.225	-.170	.205	.198	.254	1	.000	.150	.240	.553*	-.130
	Sig. (2-tailed)	.340	.475	.385	.403	.281		1.000	.527	.309	.011	.584
	N	20	20	20	20	20	20	20	20	20	20	20
R7	Pearson Correlation	-.138	.202	.045	.083	.000	.000	1	.108	.252	.045	.106
	Sig. (2-tailed)	.560	.394	.849	.727	1.000	1.000		.651	.284	.852	.656
	N	20	20	20	20	20	20	20	20	20	20	20
T1	Pearson Correlation	-.031	-.100	-.511*	.055	.345	.150	.108	1	.734**	.192	-.443
	Sig. (2-tailed)	.898	.674	.021	.818	.136	.527	.651		.000	.418	.051
	N	20	20	20	20	20	20	20	20	20	20	20
T2	Pearson Correlation	.225	.006	-.163	.053	.254	.240	.252	.734**	1	.501*	-.426
	Sig. (2-tailed)	.340	.980	.492	.824	.281	.309	.284	.000		.024	.061
	N	20	20	20	20	20	20	20	20	20	20	20
T3	Pearson Correlation	.350	-.067	.293	-.183	-.061	.553*	.045	.192	.501*	1	-.187
	Sig. (2-tailed)	.130	.778	.210	.439	.798	.011	.852	.418	.024		.430
	N	20	20	20	20	20	20	20	20	20	20	20
T5	Pearson Correlation	-.078	.408	.210	-.215	-.360	-.130	.106	-.443	-.426	-.187	1
	Sig. (2-tailed)	.742	.074	.375	.362	.119	.584	.656	.051	.061	.430	
	N	20	20	20	20	20	20	20	20	20	20	20

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Appendix O: Correlation Matrix (Form F)

		R1	R2	R3	R4	R5	R6	R7	T1	T2	T3	T5
R1	Pearson Correlation	1	.524*	.518*	.400	.424	.276	.061	-.347	-.406	.267	-.021
	Sig. (2-tailed)		.018	.019	.081	.063	.239	.797	.134	.076	.255	.929
	N	20	20	20	20	20	20	20	20	20	20	20
R2	Pearson Correlation	.524*	1	.393	.107	.173	.516*	.200	-.177	-.187	.370	.094
	Sig. (2-tailed)	.018		.086	.653	.467	.020	.398	.456	.429	.109	.692
	N	20	20	20	20	20	20	20	20	20	20	20
R3	Pearson Correlation	.518*	.393	1	-.083	-.146	.000	.222	-.419	-.686**	.387	.065
	Sig. (2-tailed)	.019	.086		.729	.539	1.000	.347	.066	.001	.092	.786
	N	20	20	20	20	20	20	20	20	20	20	20
R4	Pearson Correlation	.400	.107	-.083	1	.750**	.595**	.196	-.173	.097	.021	.174
	Sig. (2-tailed)	.081	.653	.729		.000	.006	.407	.465	.683	.929	.463
	N	20	20	20	20	20	20	20	20	20	20	20
R5	Pearson Correlation	.424	.173	-.146	.750**	1	.631**	.104	.037	.057	.113	-.040
	Sig. (2-tailed)	.063	.467	.539	.000		.003	.663	.878	.810	.635	.866
	N	20	20	20	20	20	20	20	20	20	20	20
R6	Pearson Correlation	.276	.516*	.000	.595**	.631**	1	.189	.100	-.052	.515*	.252
	Sig. (2-tailed)	.239	.020	1.000	.006	.003		.424	.674	.827	.020	.284
	N	20	20	20	20	20	20	20	20	20	20	20
R7	Pearson Correlation	.061	.200	.222	.196	.104	.189	1	.037	.000	.295	-.281
	Sig. (2-tailed)	.797	.398	.347	.407	.663	.424		.876	1.000	.207	.231
	N	20	20	20	20	20	20	20	20	20	20	20
T1	Pearson Correlation	-.347	-.177	-.419	-.173	.037	.100	.037	1	.411	-.023	.132
	Sig. (2-tailed)	.134	.456	.066	.465	.878	.674	.876		.072	.923	.578
	N	20	20	20	20	20	20	20	20	20	20	20
T2	Pearson Correlation	-.406	-.187	-.686**	.097	.057	-.052	.000	.411	1	-.238	-.155
	Sig. (2-tailed)	.076	.429	.001	.683	.810	.827	1.000	.072		.311	.514
	N	20	20	20	20	20	20	20	20	20	20	20
T3	Pearson Correlation	.267	.370	.387	.021	.113	.515*	.295	-.023	-.238	1	-.256
	Sig. (2-tailed)	.255	.109	.092	.929	.635	.020	.207	.923	.311		.275
	N	20	20	20	20	20	20	20	20	20	20	20
T5	Pearson Correlation	-.021	.094	.065	.174	-.040	.252	-.281	.132	-.155	-.256	1
	Sig. (2-tailed)	.929	.692	.786	.463	.866	.284	.231	.578	.514	.275	
	N	20	20	20	20	20	20	20	20	20	20	20

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).